DOCUMENT RESUME

ED 152 426 PS 009 849

AUTHOR Haney, Walt
TITLE The Follow Through Planned Variation Experiment. Volume V: A Technical History of the National Follow Through Evaluation.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
PUB DATE Aug 77
CONTRACT OEC-0-74-0394
NOTE 365p.; The chart and tables in this document may be marginally legible due to small print size; For related documents, see ED 094 890-891, ED 108 768-769, ED 148 488-493, and PS 009 850

EDRS PRICE MF-$0.83 HC-$19.41 Plus Postage.
DESCRIPTORS *Academic Achievement; Compensatory Education; *Disadvantaged Youth; *Evaluation Methods; Historical Reviews; Longitudinal Studies; *Primary Education; Program Effectiveness; *Program Evaluation
IDENTIFIERS *Planned Variation Program; *Project Follow Through

ABSTRACT

This report presents a technical history of the national evaluation study of the Follow Through Planned Variation Model. The last in a five volume series dealing with Follow Through evaluation activities, this volume is devoted to examining the major issues confronting the evaluation and the attempts made to resolve these issues. Chapter I provides an overview of the history of the Follow Through program and the national evaluation, delineating the changing relationship between the two and assessing the effects of this changing relationship on each. Chapter II shows how the national evaluation has narrowed to focus mainly on one sort of question -- about a subset of the Follow Through planned variations -- which represents but one of the conceptions of what is most important about Follow Through. Chapter III recounts how major national Follow Through evaluation studies have gone about answering the main questions asked, and how they have employed the treatment-control experimental paradigm to compare the effects of the Follow Through Planned Variation treatments. Chapter IV discusses how the evaluation has handled analyses of parent interview and teacher questionnaire data. Chapter V presents a summary of the analysis report and offers two types of conclusions from it: an overview of the substantive findings of the national evaluation and a commentary on the accomplishments of the evaluation. (Author/JMB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED152426

PS 009849

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

THE FOLLOW THROUGH PLANNED VARIATION EXPERIMENT

Volume V
The Follow Through Evaluation: A Technical History

DISCRIMINATION PROHIBITED. No person in the United States shall, on the ground of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving Federal financial assistance, or be so treated on the basis of sex under most education programs or activities receiving Federal assistance.

THE
FOLLOW
THROUGH
PLANNED
VARIATION
EXPERIMENT

PS 009849

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE
Joseph A. Califano, *Secretary*
Mary Berry, *Assistant Secretary for Education*
Office of Education
Ernest Boyer, *Commissioner*

THE FOLLOW THROUGH PLANNED VARIATION EXPERIMENT

Volume V:

A Technical History of the National Follow Through Evaluation

By

Walt Haney

August 1977

The Huron Institute
123 Mt. Auburn St.
Cambridge, MA 02138

(617) 661-9285

Prepared for the
Office of Planning, Budgeting and Evaluation
United States Office of Education
Department of Health, Education and Welfare
Contract No. OEC-0-74-0394

USOE Project Officers:
Gerald Burns - 1976-77
Mary Kennedy - 1974-76

Table of Contents

Preface i
Acknowledgements ii

INTRODUCTION 1

I. THE FOLLOW THROUGH PROGRAM AND THE NATIONAL EVALUATION 7

Inception and Planning 7
Pilot Phase and the Pittsburgh Evaluation 12
Reincarnation as a Planned Variation Experiment 13
The Stanford Research Institute Evaluation Effort 23
Revamping the Evaluation and the Shift to Abt 44
Days of Reckoning and the Second Phase at Abt 67

II. FOLLOW THROUGH AND THE SCOPE OF THE NATIONAL EVALUATION 82

What is Follow Through? 84
Evaluation and the Scope of the Program 93
Evaluation and the Scope of Planned Variation 100
Focus of Inquiry 108
  - What is the Follow Through Program? 108
  - Does the Program Work? 110
  - How Does the Program Work? 111
  - How Much Does the Program Cost? 113

III. ISSUES IN THE ANALYSIS OF STUDENT OUTCOME DATA 117

Selection and Assignment 122
  - The Lure of Randomization 123
  - Selection of Program Treatment 127
  - Selection and Assignment of Projects 129
  - Selection of Schools and Children 131
  - Selection of Evaluation Sample 136
  - Selection of Control Groups 138

Units and Levels of Analysis 142

Follow Through Treatments: Specification and Verification 151
  - Implementation in the Impact Studies 152
  - Classroom Observation Approaches to Assessing Implementation 157
  - Control Groups and Diffusion of Treatment 165

Duration of the Treatment and the Evaluation 170
  - Start of the Follow Through Treatment and Time of Pre-test 171
  - End of the Follow Through Treatment and Evaluation 174

Standards of Comparison and Outcome Measures 179
  - Early Approaches to Follow Through Instrumentation 179
  - Selection of a Standard Test Battery 186
  - Continuing Instrument Development Efforts 194

Attrition 202
  - Defining the Follow Through Attrition Problem 203
  - Early Handling of the Attrition Issue 205
  - The Attrition Issue in the First Phase of Abt Analyses 210
  - Feeling Better But Looking Worse 213

ANCOVA: Angel of Salvation or Temptress and Deluder 219
  - The Need for ANCOVA 220
  - Experimental Subjects and Treatments 223
  - Selection of Covariates 229
  - ANCOVA Assumptions 239
  - ANCOVA: Summing Up 250

IV. ANALYSES OF PARENT INTERVIEW AND TEACHER QUESTIONNAIRE DATA 257

Evaluation of Parent Interview Data 258

Evaluation of Teacher and Teacher Aide Questionnaire Data 271

Summing Up 280

V. SUMMARY AND CONCLUSIONS 290

Follow Through Programs and the Planned Variation Evaluation 291

Scope of the Planned Variation Evaluation 295

Analysis of Student Outcome Data 298

Analyses of Parent Interview and Teacher Questionnaire Data 304

Overview on Findings 307

Accomplishments of the National Follow Through Evaluation 333

Attachments

Glossary of Acronyms 344

Chronology of the Follow Through Program and Evaluation and Related Events 346

Tables and Figures

Page

Table 1: Follow Through Projects, Enrollments, and Funding 21

Table 2: Major National Follow Through Evaluation Studies [illegible]

Table 3: Focus of Follow Through Evaluation Reports 96

Table 4: Sponsors Included in Major Evaluation Reports 102

Table 5: Projects Included in Major Evaluation Reports 103

Table 6: Overall Comparison of Income and Preschool Experience for Samples Analyzed in Abt IV Report 141

Table 7: List of Critical Variables Selected by SRI and Sponsors 161

Table 8: Sponsor Interests According to Measurement Areas 184

Table 9: Metropolitan Achievement Tests Used in the National Follow Through Evaluation: By Version, Subtest and Content 188

Table 10: Levels of Effects Estimates and Subjects or Units of Analysis in the ANCOVA Used in the Follow Through Evaluation 224

Table 11: Parent Interviews: Follow Through and Non-Follow Through Parents (and Projects) 261

Table 12: Analyses of Parent Interview Data 268

Table 13: Analyses of Teacher Questionnaire Data 278

Table 14: Findings of the Abt Volume IV Evaluation Report 308

Table 15: Summary of the Findings from the RMC Cost Analysis of Follow Through Projects 329

Figure 1: Geographical Distribution of Follow Through Projects 88

Figure 2: Follow Through Sponsors 90-92

Figure 3: Follow Through Cohorts and Streams 94

Figure 4: Time and Cohort of Data Included in Major National Reports 105

Figure 5: The Randomized Controlled Experiment Ideal 118

Figure 6: Quasi-experimental Reality in Follow Through 120

Figure 7: The Hierarchy of Units and Levels in Follow Through 144

PREFACE

Over the span of the Follow Through Planned Variation Experiment an enormous amount of information has been produced. Much of it has been used in the preparation of this publication. A series of volumes has been assembled to provide a comprehensive overview of the research and evaluation activities that have taken place.

Volume I, prepared by the U.S. Office of Education, contains a synopsis of the key elements of the longitudinal evaluation and a synthesis of findings. Specifically, Volume I examines the central question posed by the Follow Through research and evaluation: In what ways, to what extent, and at what cost did the various Follow Through approaches to early childhood education affect the learning and behavioral outcomes of disadvantaged children?

Volume II, a four-part set prepared by Abt Associates Inc., contains the complete presentation and treatment of the analysis conducted as part of the national evaluation which began data collection in school years 1970-71 and 1971-72. Part A contains a description of the study, the educational approaches examined, a discussion of the analytic strategies and methods of presenting results, along with a summary of findings. Part B contains a comprehensive discussion of the results. Part C contains additional analyses produced after the work on Parts A and B had been completed. These analyses complement those in Parts A and B and contain the most complete statistical analyses for all cohort streams included in the evaluation. Part C Appendix contains many detailed tables of results and supporting material.

Volume III, prepared primarily by the model sponsors in conjunction with the U.S. Office of Education, describes the models and presents, where available, additional evaluative data collected and analyzed by the sponsors.

Volume IV, prepared by RMC Inc., contains cost data on twenty of the models.

Volume V, prepared by the Huron Institute, is a technical history of the national evaluation. Major issues confronting the evaluation and the attempts made to resolve them are examined in some detail. Part V Appendix contains summaries of five earlier Follow Through evaluation reports not published as part of this series and an annotated bibliography of other evaluation reports funded by USOE since the inception of the program.

ACKNOWLEDGEMENTS

Both the Follow Through program and evaluation have been so large, complex and controversial that no account of them, technical history or otherwise, can do full justice to all aspects of the Follow Through experiment and to the many people who have invested years of their lives and scores of their ideas in it. So this history may give short shrift some aspects of the program and several facets of the evaluation. Whatever success the report does achieve in illuminating the Follow Through experiment owes much to the generosity of many people who reviewed draft versions of the technical history. Here I would like to thank several people whose suggestions proved particularly helpful. Dean Robert Egbert enlarged my understanding of several aspects of Follow Through's early history. Richard Marciano gave me especially spirited comments on other historical issues. Hal Freeman's criticisms enabled me to describe several episodes with more accuracy than otherwise would have been possible. Thanks are owed also to Andy Anderson, Geoffrey Bock, Linda Stebbins and others at Abt Associates; to Elizabeth Gilkeson and Richard Feldman of Bank Street College; to David Weikart of the High/Scope Foundation; to Grace Hilliard and Jeanne McDonald of the Education Development Center; to Allan Abrams and David Berliner of the Far West Laboratory; to Johnnie Follins and Jean Humphrey of Clark College; to Wes Becker of the University of Oregon; to Charlie Smock of the University of Georgia; to Teresa Middleton and John Emrick of the Stanford Research Institute; to Jim DiCostanzo of the University of Pittsburgh; to Joseph Fillerup of the University of Arizona; to Preston Kronkosky of the Southwest Educational Development Lab, to Judith Burnes and Garry McDaniels of the U.S. Office of Education (OE) and to Lynn Weis of the University of Kansas.

Appreciation goes too to officials in OE who have had responsibility for the Follow Through program and evaluation. Rosemary Wilson, Director of the Follow Through program and Fred Bresnick both answered numerous questions. Eugene Tucker, Mary Kennedy, David Iwamoto, and Jerry Burns, who have long been involved in the Follow Through evaluation provided a variety of helpful suggestions for improving the technical history. In particular Jerry Burns, who has been project officer on the Huron Follow Through contract for the last year is owed special thanks. He has not only provided numerous ideas for improving the technical history, but has also given me unfailing support in digging out details of the evaluation and in tracking down fugitive documents from it.

Finally, colleagues here at the Huron Institute--David Cohen, Anthony Bryk, Mike Garet, John Villaume and Herbert Weisberg--have contributed in numerous ways to improving this report. And Marcia Tress, executive secretary to the Huron Follow Through project, deserves special thanks for her patience and diligence in helping to prepare the report.

All of these people are, of course, exempt from responsibility for any errors of fact or interpretation in the technical history. And for this report, such a disclaimer merits more than pro forma mention. The history recounts not just the nuts and bolts of the Follow Through evaluation, but also many of the crises and controversies which have buffeted this unique experiment. Parties to these controversies may find that this recounting fails to fully describe some of the issues at stake or the roles of people involved in them. Any such shortcomings, I truly regret. I hope only that this imperfect recounting will be judged preferable to pretending that the uncertainties and conflict of opinion over the Follow Through experiment did not exist.

W.H.
Cambridge, Mass.
August 1977.

> Follow Through, initiated in 1967, is a project for disadvantaged children from kindergarten through third grade. It is intended to be a comprehensive project offering educational, medical and dental, nutritional, social, and psychological services to children previously enrolled in Head Start. Follow Through uses a strategy of "planned variation" in approaches to early elementary education, and 20 different approaches are now being implemented in Follow Through sites throughout the nation. The Follow Through project represents a major test of the viability of exporting models, some of which have been effective in experimental sites, to other locations to be implemented by individuals other than the initiators.
>
> Sheldon White, et al, Federal Programs for Young Children: Review and Recommendations, 1973[1]

The Follow Through program has been a unique undertaking. As a planned variation experiment, it has attempted to try out different approaches to early education in different places at the same time. As a "major test" of alternative educational models, Follow Through has fostered high hopes for what would be learned from the endeavor. What difference will it make in a disadvantaged child's educational development if he or she is taught in the early elementary grades in classrooms using an open education model of instruction? Will it make any difference if the child is taught using an approach based on stimulus-response behavioral theory? Can a child's learning be enhanced by individualizing instruction in the classroom? Will children's concepts of themselves be any different if they study in open or in highly structured settings?

---

[1] Sheldon White, Mary Carol Day, Phyllis K. Freeman, Stephen A. Hantman and Katherine P. Messenger, Federal Programs for Young Children: Review and Recommendations, Vol. II: Review of Evaluation Data for Federally Sponsored Projects for Children. (Washington, D.C.: U.S. GPO, 1973), p. 83.

In Follow Through, vast resources have been devoted to answering such questions. Altogether around 50 million dollars have been invested in the national Follow Through evaluation, and around a half-billion dollars in the program itself. Thousands of people have been employed simply to test Follow Through children and hundreds more to analyze the evaluation data. Reports on the program occupy several yards of shelf space and fill dozens of file drawers.

The purpose of this report is to provide a guide through the realms of Follow Through evaluation reports and in a preliminary way to digest and summarize what they say and point out what they do not say. The report has five chapters. The first provides an overview of the history of the Follow Through program and the national evaluation. It recounts some of the twists and turns in the relationship between program and the national evaluation; how Follow Through began as a service program; how it was partially transformed into a planned variation experiment; how the clouded conception of the program led to confusion for the evaluation; and how the evaluation, when firmly latched to the planned variation conception of Follow Through was still troubled by numerous organizational and technical issues.

Chapter II looks more closely at the fit between the Follow Through program and the evaluation. The chapter begins with a definition of the program and an outline of the program's dimensions. Against this backdrop the scope of the evaluation is then viewed from three different angles:

in terms of a set of evaluation criteria suggested as appropriate to the program; in light of the range of planned variations in the program; and finally, in view of the types of evaluation questions that might have been asked about it. In sum, the chapter shows how the national evaluation has narrowed to focus mainly on one sort of question, about a subset of the Follow Through planned variations, which represents but one of the conceptions of what is most important about Follow Through.

Chapter III recounts how major national Follow Through evaluation studies have gone about answering the main questions asked, and how they have employed the treatment-control experimental paradigm to compare the effects of the Follow Through planned variation "treatments." The national Follow Through evaluation studies are reviewed in terms of how they have handled a series of analytical problems troubling the application of the paradigm. Specifically, the chapter treats each of the following: selection and assignment, units and levels of analysis, definition and verification of "treatments," the duration of the treatment and the evaluation, outcome measures, attrition and the use of analysis of covariance (ANCOVA).

Chapter IV recounts how the evaluation has handled analyses of parent interview and teacher questionnaire data. It describes how efforts to assess the opinions and attitudes of parents and teachers involved in Follow Through have been troubled by lack of continuity and ambiguity of purpose. The chapter relates how the analyses of parent interview and teacher questionnaire data have varied and changed through the years until finally in the last of the major evaluation reports on Follow Through, they have faded into near oblivion.

Chapter V presents a summary of the analysis report and offers some overall conclusions from it. Conclusions are divided into two parts. The first provides an overview of the substantive findings of the national evaluation, and the second offers some comments on the accomplishments of the national Follow Through evaluation.

Appended to the history are several pages designed to help the reader who is unfamiliar with the Follow Through program and its history. First is a glossary of acronyms, dozens of which seem inevitably to seep into any description of government programs and organizations. Second is a simple chronology of events related to the Follow Through program and evaluation. The chronology hopefully will aid the reader to follow the overall flow of events, evaluations and reports described in the technical history.

The technical history has no bibliography attached to it. A bibliography was prepared, but since it turned out to be quite bulky, and of some interest in and of itself, it is being issued as a separate appendix.[2] Also it should be noted that many of the documents drawn upon in this review never were published or even reached final report form. They exist only as draft reports or papers for a variety of reasons; for example, because personnel changed, program or evaluation priorities shifted, or because separate drafts were merged in final report form. Nevertheless, a variety of these draft materials have been used as sources because they often contain valuable information not available elsewhere.

Among the fugitive literature on Follow Through is a previous version of this report.[3] This final report differs from it in several important respects. First, the draft Analysis of Interim Follow Through Evaluation

---

[2] John Villaume and Walt Haney, Volume V Appendix: Analysis of Interim Follow Through Evaluation Reports, in the current series. Hereafter this document is called simply the bibliography.

[3] Walt Haney, Draft Analysis of Interim Follow Through Evaluation Reports (Cambridge, MA: Huron Institute, September 1976).

Reports traced the history of the national FT evaluation only through 1975. The current report traces the national evaluation through April 1977. Second, the current report contains a discussion of issues in the analysis of children's outcome scores in chapter III which is much expanded over the corresponding chapter in the draft report. Third, Chapter IV in the current version of the analysis report is completely new. Finally, the summary and conclusions chapter has been almost entirely rewritten in light of foregoing changes.

Before proceeding to the narrative history of Follow Through and its evaluation, a brief explanation concerning the limitations of this paper is in order. First, in focusing upon national-level evaluations of the Follow Through program, this paper does not cover the multitude of evaluation reports by Follow Through sponsors and local education agencies.[4]

Regarding sources it should be noted that this report is based mainly on written source materials. Although conversations and interviews with various people involved with Follow Through have often expanded my understanding of aspects of the Follow Through Experiment, I have not been able to interview systematically most of the people involved in planning or administration of the Follow Through program and evaluation. Hence, this report is based largely on the written rather than the oral history of Follow Through.

---

[4] Sponsor reports for the 1972-73 school year were reviewed in Gregory Jackson's Review of Follow Through Sponsor and Local Data. (Cambridge, MA: Huron Institute, 1974), and sponsor evaluation reports for the 1970-71 school year were reviewed in Robert Hardy and Robert Huebner, Evaluation of Follow Through Programs by Follow Through Sponsors 1970-71. A summary Analysis and Review. (Bureau of Elementary and Secondary Education), OE, HEW: Washington, D.C., December 1971). Both of these documents are described briefly in the bibliography.

Also, it should be pointed out that different staff members at the Huron Institute have been involved in the Follow Through evaluation for approximately five years. Huron has had several contracts with the Office of Education concerning the Follow Through evaluation dating back to 1972. In that year Huron staff members played an active part in redesigning the national Follow Through evaluation. Since 1972, Huron has provided technical assistance to other actors in the Follow Through evaluation story and has conducted a variety of special studies for the Follow Through Evaluation Branch (these are described in the bibliography). Personally, I have been previously involved in Huron's Follow Through project in two ways--in preparing a report on units-of-analysis in the Follow Through evaluation and in providing technical assistance on other aspects of the evaluation. This experience has affected my thinking,as is evident in subsequent pages. Nevertheless, my own previous involvement in the Follow Through evaluation has not, I trust, tinged my vision with an over-rosy hue. Likewise, I hope that familiarity has not bred overmuch distrust. On balance, I have attempted to describe the Follow Through evaluation and documents from it fairly and as best as I understand them.

I. THE FOLLOW THROUGH PROGRAM AND THE NATIONAL EVALUATION

Before attacking the stacks of reports and evaluations on Project Follow Through, some perspective on the whole project is essential. Otherwise, it is impossible to make sense of the way in which evaluation questions were dropped or added along the way. This account of the Follow Through program and evaluation is necessarily brief. It draws heavily on three much more substantial accounts[1], on other evaluation reports and especially for later periods, on memoranda and correspondence among participants.

Inception and Planning

In his State of the Union address to Congress in January 1967, President Lyndon Johnson proposed that "we should strengthen the Head Start program, begin it for children 3 years old, and maintain its educational momentum by following through in the early years of school."[2] In his Special Message on Children and Youth in February, the President asked Congress "to preserve the hope and opportunity of Head Start by a 'followthrough' program," adding that "to fulfill the rights of America's children to equality of educational opportunity, the benefits of Head Start must be carried to the early grades."[3]

---

[1] Stanford Research Institute, Draft, Appendix F. Administrative History of the Follow Through Evaluation 1968-1972 (Menlo Park, CA: Stanford Research Institute, August, 1972); Gilbert Krulee, An Organizational Analysis of Project Follow Through, Final Report (Evanston, IL: Northwestern University, April, 1973); Richard Frazier Elmore, Follow Through Decisionmaking in a Large-Scale Social Experiment, doctoral dissertation Harvard Graduate School of Education, 1976.

[2] U.S. Congress, House, State of the Union Message by President Lyndon B. Johnson, 90th Congress, 1st session, January 10, 1967, Congressional Record Vol. 113, pt. 1, 37. Quoted in R. Elmore, Follow Through: Decision-making in a Large-Scale Social Experiment, pp. 111-12.

[3] New York Times, February 9, 1967. Also U.S. Congress, House, Message on Children and Youth by President Lyndon B. Johnson, 90th Cong., 1st sess., February 8, 1967, Congressional Record, Vol. 113, pt. 1, 2882.

The Head Start program had been authorized under Title II of the Economic Opportunity Act of 1964, and begun in the summer of 1965. The program quickly became the most popular of the Great Society programs, and by fall 1967 more than two million children nationwide had attended Head Start.[4] Early evaluations showed that Head Start had benefited children in several respects, but follow-up evaluations suggested that Head Start graduates tended to level off, allowing children who didn't attend Head Start to catch up early in their school experience.[5] The relative loss of preschool gains was variously called fadeout, wash-out, or dissipation, but by whatever name, the phenomenon gained considerable attention. In November 1966, in a speech in Milwaukee, Director of the Office of Economic Opportunity (OEO), Sargent Shriver, declared that "The readiness and receptivity gained in Head Start has been crushed by the broken promises of first grade."[6]

The most widely cited of the follow-up studies on Head Start children was an investigation by Max Wolff and Annie Stein of Yeshiva University.[7]

M. S. Smith and J. S. Bissell, "Report Analysis. The Impact of Head Start", Harvard Educational Review, 40:1 (1970), 51-104.

[5] E. Grotberg, "Review of Research 1965 to 1969," Research and Evaluation Office, Project Head Start, Office of Economic Opportunity, June 1969.

[6] Quoted in Robert Egbert, "Follow Through," unpublished manuscript, n.d. [ca. 1971], p. 2.

[7] Max Wolff and Annie Stein, Six Months Later: A Comparison of Children Who Had Head Start, Summer 1965, with Their Classmates in Kindergarten (New York: Ferkauf Graduate School of Education, Yeshiva University, August 1966, ERIC No. ED015025.)

The Wolff and Stein study compared kindergarten children who had participated in the Head Start program with their classmates who had not. The children had attended three New York City Head Start centers which had been rated "very good to excellent" by Head Start supervisors. At the time of the study, the Head Start graduates were attending kindergarten in four New York public elementary schools. Wolff and Stein investigated the kindergarten status of the Head Start graduates in three ways; by interviewing parents, interviewing teachers, and testing children. They found that parents were unanimous in their enthusiasm for Head Start, that teachers differed in their evaluation of the influence of Head Start and that "There was no significant difference between the scores of Head Start children and their classmates in kindergarten who did not have Head Start as measured by the [Caldwell] Preschool Inventory six to eight months after the summer Head Start experience."[8] On the basis of these findings, Wolff and Stein suggested that there was a paucity of direct learning in the kindergarten curriculum and that the effects of Head Start were dissipated during the first year of elementary school. This inference was questionable, however, since the Wolff and Stein study had not shown that there were any test score gains attributable to Head Start in the first place.[9] Yet despite methodological weaknesses, the Wolff and Stein study gained considerable notoriety. Writing in 1968 about the origins of Follow Through, the first Follow Through director quoted directly from the Wolff and Stein study, and later recounted that "the decision to request a Follow Through program was stimulated by [that] single study."[1] So, building on the popularity of Head Start, and with seeming support from empirical investigation, as well as powerful political backing, the proposal for Follow Through quickly moved toward reality.

---

[8] These conclusions are recounted from an article summarizing the study. Max Wolff and Annie Stein, "Head Start Six Months Later," Phi Delta Kappan, March 1967, pp. 350-51.

[9] For critiques of the Wolff and Stein study see Urie Bronfenbrenner, "Memorandum: Comments on Wolff and Stein Study," January 1967 (ED 015 029), and Edmund Gordon, "Remarks on the Max Wolff Report," undated (ED 015 030).

In fact, long before Johnson's proposal for Follow Through had passed through Congress, work began on the new program. Follow Through was to be authorized under the Economic Opportunity Act and, therefore, would be the official responsibility of the OEO, but one of the first planning decisions was for the OEO to delegate responsibility for Follow Through to the Office of Education (OE). Official authorization to move Follow Through from OEO to OE was not signed by President Johnson until June of 1967, but early in the year the OE organized a National Advisory Committee for Follow Through.

---

[1] R. Egbert, "Follow Through: Fulfilling the Promise of Head Start" in J. Hellmuth, Disadvantaged Child 2: Head Start and Early Intervention (New York: Brunner/Mazel, 1968), and R. Egbert, "Planned Variation in Follow Through," unpublished manuscript, p. 2. This paper was presented to the Brookings Institution's Panel on Social Experimentation (Washington, D.C.: April 1973). A number of papers presented to this panel were later published in A.M. Rivlin and P.M. Timpane (eds.), Planned Variation in Education: Should We Give Up or Try Harder? (Washington, D.C.: Brookings Institution, 1975). Unfortunately, Egbert's paper was not included in this volume, but is available in Hearings of the Committee on Education and Labor, House of Representatives, February 21, 1974, pp. 121-31.

Elmore suggests that the influence of the Wolff and Stein study was somewhat ambiguous. "The mechanism by which policymakers found support for Follow Through in the Wolff and Stein study is a familiar one: Having already decided on other grounds--logical and political--that Follow Through was desirable and necessary, they were prepared to read support for their position into the study's ambiguous and flawed conclusions." (Elmore, Follow Through: Decisionmaking, p. 125.)

chaired by Gordon Klopf of the Bank Street College of Education. The advisory group met several times in Washington during the first few months of 1967, and produced a series of task force reports which provided a basis for much of a later draft program manual for Follow Through.[2] In mid-June a panel was convened to select sites for the pilot phase of the program. On July 2, Commissioner of Education Harold Howe and OEO Director Sargent Shriver announced the selection of 30 pilot sites and predicted full-scale expansion of the program for the 1968-69 school year. In late August, Robert Egbert, who had been a member of the Advisory Committee and Chairman of its Task Force on Research and Evaluation, was appointed director of the program.[3]

The planning phase came to a close at the end of summer 1967 with 30 projects funded at $85,000 each to begin that fall. So in September 1967 the follow-through on Head Start had become a reality.[4] In the words of the legislative proposal sent to Congress by the Johnson administration the program

> ...known as 'Follow Through' focused primarily on children in kindergarten or elementary school who were previously enrolled in Headstart or similar programs and designed to provide comprehensive services and parent participation activities as described in [the Head Start] authorization which the Director [of OEO] finds will aid in the continued development of children to their full potential.[5]

---

[2] Bank Street College of Education; Follow Through Task Force Reports, New York: Bank Street, undated [1967].

[3] Egbert was preceded by two acting directors: first Alexander Plante and then Eugene Walsh.

[4] Krulee, p. 24. Ten more sites were selected in late summer, to join the program in the second semester of the 1967-68 school year.

[5] Section 222 (a) (2), Economic Opportunity Act of 1964, P. L. 90-22 as amended.

Pilot Phase and the Pittsburgh Evaluation

Although plans for Follow Through called for comprehensive services and parent participation, the exact manner in which Follow Through was to "aid in the continued development of children to their full potential" was not spelled out in much detail. Nevertheless, a contract was let during the summer of 1967 to the University of Pittsburgh to conduct a national evaluation of Follow Through. Plans for the Pittsburgh evaluation were ambitious, calling for evaluation "at three levels, i.e., product, process and cost."[6]

The Pittsburgh evaluation effort entailed a wide variety of data gathering; pre- and post-tests on three cognitive instruments, one non-cognitive instrument, a medical-dental questionnaire and a "teachers Health Observation Form." In addition, the Pittsburgh evaluators planned to investigate the processes of the intervention, since:

> Answering the question, "What changes took place in the Follow Through population over the period of the program" is essentially meaningless in terms of the evaluative function unless one can also answer the question "What actually occurred to produce the changes?"[7]

The Pittsburgh evaluation fell completely short of its promise. Despite an eight-month extension, the Pittsburgh evaluators were unable to complete their plans for analyzing post-test data. As a result, the Pittsburgh analyses consisted mainly of comparisons of pre-test scores for Follow Through children with those of children not in the program. After such a fruitless effort, Kurfeerst and Stephens advised:

---

[6] M. Kurfeerst and T. M. Stephens, The National Evaluation of Project Follow Through: 1967-1968 (Pittsburgh, PA: University of Pittsburgh, 1969, ERIC No. 078 923), p. 1.

[7] Kurfeerst and Stephens, pp. 69-70.

> It cannot be overemphasized that an agreed-upon procedure for monitoring the orderly progress of any phase of the project (including evaluation) should be implemented.[8]

A less-charitable view of the Pittsburgh evaluation was later advanced by two Nader's Raiders, Daniel Guttman and Barry Willner. They observed that the $146,000 evaluation by Pittsburgh was "useless" and commented:

> ...a July 1967 memo from the Deputy Commissioner of Education explained that the entire contract had been a mistake. The award was 'not based upon the quality of the Pittsburgh proposal, but primarily upon the basis of an oral commitment made to Pittsburgh in violation of internal proceedings.'[9]

The Pittsburgh evaluation report was completed in May 1969, but understandably never was widely circulated. It was not an auspicious beginning for the Follow Through evaluation. Yet in the meantime, more than a year before the Pittsburgh evaluation report was finished, the whole course of the Follow Through program had shifted and the evaluation was to begin anew from scratch.

Reincarnation as a Planned Variation Experiment

Follow Through was originally conceived as a broad-scale social action program, like a sort of Head Start continued into the early elementary grades.

---

[8] Kurfeerst and Stephens, p. 80.

[9] D. Guttman and B. Willner, The Shadow Government. The Government's Multi-Billion-Dollar Giveaway of Its Decision Making Powers to Private Management Consultants, "Experts", and Think Tanks (NY: Pantheon Books, 1976), p. 159. The Pittsburgh analysis was a 2x2 ANOVA, with factors FT vs. NFT and HS vs. NHS. The analysis was carried out on only 22 of the 29 FT sites in the evaluation. In seven sites, one or more of the four cells had too few subjects ($n < 5$) to allow analysis. The thirtieth of the original FT sites was not included in the Pittsburgh evaluation since it was "given permission to conduct its own evaluation".

The pilot phase was planned and initiated on the basis of this conception and up until the fall of 1967, Follow Through planners expected that the new programs would receive an allocation of $120 million to cover the 1967-68 pilot phase and a vastly expanded program in the following school year. But during 1967, Congress was becoming increasingly unhappy with the community action programs of OEO. So when Congress finally passed the amendments to the Economic Opportunity Act in December 1967, after several months of legislative maneuvering, the OEO budget was cut by more than $200 million.[1] As a result the Follow Through budget was slashed to $15 million. Follow Through's continued existence as a national social action program was clearly in jeopardy. As one Follow Through staff member explained, "We had to find a new identity or risk extinction..."[2]

The new identity for Follow Through was that of a planned variation experiment. The OEO budget cut was the immediate catalyst for transformation of Follow Through into a planned variation experiment, but intellectual and political stimulation for such an experiment went considerably deeper than debits and credits in the budget. In September of 1967, Special Assistant to the President, Joseph Califano, requested that the Department of Health, Education and Welfare review existing federal programs for children and ideas on child development advanced in 1966, and on the basis of its review develop legislative and administrative recommendations for health, education and welfare programs for children.

---

[1] New York Times, December 16, 1967.

[2] Elmore, Follow Through: Decisionmaking, p. 158.

In response to this request Assistant Secretary of HEW for Planning and Evaluation, William Gorham, organized the 1967 Task Force on Child Development. The panel made twenty specific recommendations altogether, but two in particular foreshadowed the reincarnation of Follow Through as a planned variation experiment:

- A major effort to use Federal funds to learn how to make education of disadvantaged children more effective.
- Use of Follow Through program to evaluate major variations in compensatory education programs in the early elementary grades.[3]

In a background section preceding these rec mmendations, the Task Force reported that

> studies have amply documented the inadequate educational performance of low-income and minority group children. These children start school at a disadvantage. As they pass through the school system they fall more and more years behind the majority in educational achievement measured by standard tests. They have high dropout rates and those who finish high school have poor chances for higher education both because of financial barriers and because of their relatively poor high school records.[4]

---

[3] [William Gorham, Chairman], Report of the Task Force of Child Development, November 7, 1967. For more detail on this task force see Elmore, FT: Decisionmaking, pp. 158-163.

[4] Report of the Task Force on Child Development, p. 18.

Summarizing the state of knowledge concerning education for disadvantaged children, the report concluded

> The one fact that stands out in all our efforts to analyze education programs is our lack of knowledge about effective ways of raising the educational attainment of students. Research techniques must be combined with experimentation and innovation to throw light on this problem, and widespread dissemination of the most effective methods must be undertaken.[5]

The panel recommended that the new FT program offered an ideal opportunity to remedy the knowledge deficit; specifically to explore the effects of a wide variety of compensatory education programs and evaluate what works under what circumstances. Apparently, as a result of its enthusiasm for the opportunities afforded by FT, the Task Force set out its recommendations on the new program in detail:

> In consultation with the Office of Economic Opportunity, the Commissioner of Education should develop a plan for Follow Through projects that systematically vary
> --curriculum
> --pupil-instructional staff ratio
> --racial and socio-economic mix
> --age and training of teachers and aides
> --type of parent involvement.
>
> At the same time, he should develop a plan of evaluation using
>
> --multiple criteria of success (achievement, self-esteem, and social competence)
> --comparable measurements in as many programs as possible.
>
> The Administration should explicitly provide budget and personnel allowances for a Follow Through staff to stimulate and develop projects consistent with these plans and to accomplish the required evaluations.
>
> The Office of Education should select all new Follow Through projects in accordance with these plans for major variation and evaluation.

---

[5] Report of the Task Force on Child Development, p. 19.

> The Administration should assure adequate funding to Follow Through, no less than $40 million in fiscal year 1968 and $80 million in fiscal year 1969, since Follow Through represents a major opportunity for learning what works for what groups of disadvantaged children in the early elementary grades and discontinuance of Follow Through risks losing for the child the benefits of his Head Start experience.[6]

There were two main problems with the Task Force recommendations for FT.[7] First, FT had already begun--not as systematic experiment, but as a service-oriented follow through on Head Start. Second, it would be an unprecedentedly difficult task for the Office of Education officials to administer an experiment which systematically varied different features of instructional program. Also it ran the risk of raising charges of unwarranted federal intervention in education.

Uncertainty over how to set up and run the planned variation experiment was greatly eased when a new feature was added to the planned variation idea--the concept of model sponsors. The notion of sponsors was the brainchild of FT Director, Robert Egbert and his Chief of Evaluation, Richard Snyder. From the point of view of administering the planned variation experiment, the idea of model sponsors was, as Richard Elmore termed it, a stroke of genius.[8] The plan was that the Office of Education would employ groups advocating different curriculum approaches for educating young children to act as consultants to local FT projects. These groups--model sponsors as they came to be called--would help implement their models in different communities around the country. Then, went the theory, the results of these planned variations could be compared in order to determine which curriculum models were most effective in producing which kinds of educational outcomes for which kinds of children. The first FT Director, Robert Egbert, later summed up the shift in identity as follows:

---

[6] Report of the Task Force on Child Development, pp. 20-21.

[7] It is unclear exactly how much influence the Task Force report had on the decision to use a planned variation approach in FT. Apparently much of the support for pushing FT in the planned variation direction came from the staff of HEW Assistant Secretary for Planning and Evaluation (ASPE) Gorham, who also provided staff support to the Child Development Task Force. See Elmore, FT: Decisionmaking, pp. 163-165. Milbrey McLaughlin suggests that the failure of the TEMPO evaluation of Title I in particular prompted ASPE to support planned variation in FT. Milbrey Wallin McLaughlin, Evaluation and Reform The Elementary and Secondary Education Act of 1965, Title I. (Cambridge, MA: Ballinger, 1975) p. 41.

> With such limited funds it seemed sensible to change Follow Through's primary purpose from 'service to children' to 'finding out what works'...Follow Through now focused its attention on developing, examining, and refining alternative approaches to the education and development of young disadvantaged children.[9]

Implicit in Egbert's words are the shadows of what became a major dilemma for FT. The different aims mentioned by the FT Director -- service to children, finding out what works, and developing and refining alternatives -- all seem worthwhile and are obviously related. It was not so obvious, however, that they represented alternative conceptions of program purpose, and as such, were to have major consequences for the FT evaluation. But at the start, the implications of such ambiguity seemed small cause for concern.

---

[8] Elmore, FT: Decisionmaking, p. 189.

[9] Egbert, "Planned Variation in Follow Through," unpublished manuscript pp. 4-5.

At the close of 1967, the idea of a planned variation experiment was enticing, and time was short. The problems of putting the ideas into practice were immense. Egbert and Richard Snyder, began the task in December 1967. In that month the FT officials discussed their ideas for a planned variation experiment with experts in child development and began to identify promising curriculum models. In January 1968, meetings were held in Washington for prospective model sponsors. Of the roughly two dozen groups attending these meetings, half became sponsors in the first year of planned variation. Sponsors were paired with communities in two meetings in Kansas City, Missouri in February 1968, although some communities chose not to affiliate with any sponsor. Around three-quarters of the 92 Follow Through communities for 1968-69--52 new communities in addition to the 40 pilot sites from the 1967-68 school year--were matched with sponsors.

The notion of model sponsors was brand new and so, not surprisingly, the matching of sponsors with communities was a bit chaotic. Sponsors gave brief presentations of their curriculum approaches--one sponsor later described it, "Your life's work in twenty minutes, preferably with colored slides." Community representatives then selected the sponsors they liked. As another sponsor described the process, "It was kind of a smorgasbord and we were the dishes."[1]

---

[1] The sponsors' descriptions of the Kansas City meetings came from interviews by Mary Kennedy, Fall 1975, Ypsilanti, MI.

Despite the newness of it all, the Kansas City meetings did succeed in pairing communities with thirteen different curriculum model sponsors. The sponsors varied in approach all the way from models based on theories of behaviorism--like the Behavioral Analysis Approach sponsored by the University of Kansas--to unstructured approaches like the Open Education Model sponsored by the Education Development Center.[2] For the next school year, 1969-70, eight additional sponsors were added and in 1970-71 two more sponsors joined the program. As the number of Follow Through sponsors grew, so did the number of participating communities; to the 92 in 1968-70, approximately 60 were added in 1969-70, and another twelve in 1970-71. The growth of the Follow Through Planned Variation program through these years is summarized in Table 1.

As Elmore suggests, it is easy in retrospect to discount the difficulties in getting the planned variation experiment under way. Not only was the endeavor completely novel, but it also faced political opposition even before it began. According to one observer, a high-level OE official "got wind of what was happening" just before the Kansas City meetings, and

> ...threatened to stop the whole thing by mobilizing state and local interest groups against it. He said the planned variation idea was an unwarranted intervention by the federal government in the affairs of local school districts. We managed to put the squeeze on him and shut him up. But everyone was worried that the Kansas City meeting would blow the whole thing sky high. We were very cautious about the federal control issue.[3]

---

[2] More complete description of Follow Through sponsors is given in Chapter 2.

[3] Former HEW official quoted in Elmore, FT: Decisionmaking, p. 213.

TABLE I: FOLLOW THROUGH PROJECTS, ENROLLMENTS AND FUNDING

	1967-68	1968-69	1969-70	1970-71	1971-72	1972-73	1973-74	1974-75	1975-76	1976-77
Number of projects[a]	39	92	160	178	178	173	170	169	165	164
Children enrolled from low income families[a]	2,900	15,500	37,000	60,200	78,170	84,000	81,000	78,000	76,500	75,700
Children rostered in FT[b]			46,650	71,155	90,524	101,974	69,226	89,099	-	-
Rostered projects[b]			157	177	178	174	172	171	-	-
Federal funds available (in millions of $)	3.75	11.25	32.00	70.30	69.00	63.06	50.62	52.85	55.42	59.00

---

[a] Division of Follow Through, Bureau of Elementary and Secondary Education, November 1976. A few of these figures differ from those given in previously published accounts. For example, the Comptroller General, *Follow Through: Lessons Learned From Its Evaluation and Need to Improve Its Administration* (Washington, D.C.: Comptroller General, October 1975) lists Follow Through as having 103 projects in 1968-69, 161 in 1969-70, and 78,000 children from low-income families enrolled in 1973-74. These discrepancies apparently derive from different methods of accounting.

[b] Follow Through Evaluation Project, Stanford Research Institute, July 1976. SRI rostered Follow Through children "in all grades (kindergarten, entering or non-entering First, Second, and Third) in all years *except* 1973/74 when no kindergarten or entering-first grade children were rostered." In some cases, the number of rostered projects is less than the number of total projects because one or two projects did not return roster data to SRI. Why the number of rostered projects is greater than the total in 1972-73 and 1973-74 is a puzzle.

Another problem for the planned variation experiment, in the words of Follow Through Director Egbert, was that "in the absence of any precedents, no one knew the capacity of sponsors, either individually or collectively, to work with communities. Whether a sponsor could work with two or five or ten projects in varying locations was totally unknown."[4] Another severe problem was that none of the sponsors' curriculum models was terribly well specified. At least some sponsors have candidly admitted that their models were far from ready for immediate implementation at the time they agreed to participate in Follow Through.[5] As Follow Through Director Egbert suggested, the goal of the planned variation experiment was not simply testing out what works, but also developing something that worked.

So by the fall of 1968, four major conceptions of the purpose of Follow Through had begun to shape the program. On one hand, it was a service program aimed at providing disadvantaged children with a range of educational, medical, dental, nutritional, social psychological services. Also, it was a social action program mandating parent participation and community involvement. From another perspective, it was a planned variation experiment designed to test alternative models of education. And at the same time, it was also a research and development effort designed to develop and refine alternative models. These different conceptions of Follow Through were not necessarily incompatible, or so it seemed in 1968. After all, parties holding different conceptions of Follow Through had little cause for quarrel when the very existence of Follow Through was in question.

---

[4] Egbert, "Planned Variation in FT," p. 7.

[5] Rivlin & Timpane, Planned Variation in Education, p. 15.

As long as the program was in jeopardy, the more it could be all things to all people, the more likely it would survive. The full implications of the disparate perceptions of what Follow Through was would not become clear until it came time to evaluate the program.

The Stanford Research Institute Evaluation Effort

In mid-May 1968, Dr. Richard Snyder, Chief of Follow Through Research and Evaluation, flew to California to discuss whether the Stanford Research Institute "might be interested in and capable of assisting USOE/FT in an evaluation of Follow Through."[6] Snyder also approached other research organizations about the FT evaluation. Two firms besides SRI submitted proposals for the evaluation contract and on May 31, representatives of the three research firms met in Washington with Office of Education officials to discuss the contract. In early June, SRI was selected as the evaluation contractor and invited to submit a revised proposal. The new proposal was accepted and on June 27, less than two months after preliminary proposals were requested, and just before the end of the fiscal year, the contract was officially awarded to SRI.[7]

---

[6] SRI, Appendix F: Draft Administrative History, p. F-9. The exact manner and timing of the process by which SRI was awarded the Follow Through evaluation contract are unclear. In addition to the Administrative History, sources on this topic are: Krulee, Organizational Analysis; Elmore, Follow Through: Decisionmaking; and HEW Audit Agency, Draft Report on Review of Administration of Performance Under Contracts with the Stanford Research Institute, Office of Education and Office of Child Development, December, 1972, 50 pp.

The terms of the SRI contract called for evaluation activities in five main areas: "(1) working with sponsors to specify the content and objectives of their models; (2) selection of measurement instruments to be used in the evaluation; (3) construction of an evaluation design, including selection of comparison group; (4) data collection, including training of on-site test administrators and interviewers and (5) analysis of the data collected."[8]

The evaluation contract was funded at approximately $900,000 and the SRI evaluation team was headed by Philip Sorensen and William Madow. Sorensen, the project director, was an educational psychologist, and Madow, the associate director, a well-known mathematical statistician. Both were "senior members of the SRI staff at the time."[9]

The contract award had been rushed and the scope of the contract was enormous. The rush and enormity quickly caught up with the SRI investigators. According to Krulee's account of this period, "there were too many decisions to be made and too much to be done, given the staff at hand and the pressures of time...There was a sense of crisis, of barely being able to manage and of there being too much to do..."[1]

---

[7] The Audit Agency draft report concludes that the award "was made without adequate competition and without allowing potential contractors sufficient time to prepare and submit proposals" (p. 5). The SRI Administrative History notes, without elaboration, that high-level OE consultants, including Urie Bronfenbrenner of Cornell University and Sheldon White of Harvard University, "were present during oral presentations by potential contractors in early June 1968 in Washington" (p. F-10).

[8] Unsigned and undated list of Follow Through implementation and research contracts, circa spring, 1968 (processed). Quoted in Elmore, Follow Through: Decisionmaking, p. 224.

[9] Krulee, Organizational Analysis, p. 191.

During the summer, it was decided that testing would be conducted at forty-seven sites in fall 1968. In the rush to prepare for the fall testing, the evaluators were confronted with a multitude of problems. The major questions raised at an evaluation planning meeting in early August 1968 were:

1. What set of test instruments should be included in order to provide a fair assessment of diverse sponsor goals?
2. How should project sites be selected for inclusion in the evaluation?
3. Should the evaluation be directed toward summative assessment or in formative fashion toward program development?
4. How can the education processes of the program be assessed?
5. How should institutional change be defined and evaluated?[2]

At the August 1968 meeting, none of these issues was thoroughly resolved. On the measures issue, it was decided that fall testing proceed using existing measurement instruments, but in the longer term, new instruments would be developed. The evaluation would include those "one or two projects that each sponsor considered exemplary of his approach and exclude (at least initially) projects in which difficulties in implementation were being encountered."[3]

---

[1] Krulee, Organizational Analysis, pp. 193, 202.

[2] These issues are paraphrased from SRI's Administrative History, pp. 11-12. The SRI account lists six issues. I have collapsed SRI's first two into a single question because they are closely related.

[3] SRI, Draft Administrative History, p. 11.

It was also agreed that the evaluation should be formative rather than summative during the early stages of Follow Through. Participants also apparently agreed on a strategy for obtaining process measures "through direct observation of classrooms, information about socioeconomic and other characteristics of children through information from parents and measures of teacher attitudes." The problem of defining and evaluating institutional change was "not pursued to closure."[4]

In sum, these hard issues on how to conduct the evaluation were handled in three ways: (1) they were simply left unresolved, (2) resolved only temporarily for expendiency's sake but left open for the future, or (3) answered in such vague and general terms that exactly how they were to be resolved remained unclear. All of the issues raised in August 1968, as well as the means of handling them, were to trouble the evaluation for years. At various times some issues would rise to the fore and others recede, but none of these issues, raised at the very start of the planned variation evaluation, were ever resolved to the satisfaction of most people involved in the program. One SRI staff member, in an interview with Richard Elmore, characterized the mood of discussion on these crucial evaluation issues during the first year of the evaluation as "ambient ambiguity."[5] The ambiguity remained for years to come, but the ambience was not so stable.

In October 1968, a stormy meeting in Atlanta intruded sharply on the thinking of FT planners and shifted the evaluators' efforts in a new direction. The Atlanta gathering was a general national FT meeting involving

---

[4] SRI, Draft Administrative History, p. 12.

[5] Elmore, FT: Decisionmaking, p. 251.

OE Follow Through staff, sponsors, parent representatives and others. Perhaps most influential were the general consultants who had been appointed at the suggestion of OEO to aid in local program implementation. They were not connected with any of the Follow Through model sponsors but were strongly inclined toward the social action conception of FT. At the Atlanta meeting, the general consultants charged that sponsors and SRI evaluators held a biased view of the FT program and were insensitive to the parental and community involvement aspects of the program. The general consultants also accused sponsors and evaluators, and by implication the whole program, of insensitivity to the needs of ethnic minorities. FT Director Egbert later recalled the effect of the Atlanta meeting on FT staff, sponsors, and general consultants. "This meeting followed Dr. King's assassination by only six months, and feelings were strong and powerfully expressed. For some this was the first direct contact with black anger, articulately and personally expressed. Few people left Atlanta unchanged."[6]

Moreover, neither the FT program nor the evaluation emerged from the Atlanta meetings unchanged. The main impact on the program was the addition of several new sponsors, representing minority institutions, and the revision of draft program guidelines to place more emphasis on parental involvement. The evaluation too was pulled more toward the social action conception of Follow Through. The SRI history of the evaluation recounts, "The main consequence of the confrontations in Atlanta was to make clear, sometimes dramatically and emotionally, that institutional change objectives within the adult community were viewed by many as of equal or greater importance than more immediate goals of child growth and development" (pp. 12-13). During 1968-69, SRI "began case studies in three individual project communities...to deal constructively with the importance attached to institutional change objectives" (p. 13). The SRI evaluation effort gradually expanded to include cognitive and noncognitive testing of individual children, the development of new test instruments, interviewing of Follow Through parents, as well as the assessment of institutional and community change.[7]

---

[6] Egbert, Planned Variation in Follow Through, p. 10. On the Atlanta episode, see also W. R. Rhine, "Strategies for Evaluating Follow Through" in R. M. Rippey (ed.), Studies in Transactional Evaluation (Berkeley: McCuthan, 1973) pp. 157-179.

Within the widening scope of the evaluation effort, ambiguity prevailed as to priorities. Only occasionally, at meetings bringing together interest groups holding different conceptions of Follow Through, did the ambiguity flare up into controversy. Interests clashed, for example, in an April 1969 meeting at Pajaro Dunes, California, to review SRI evaluation plans for the 1969-70 school year. Some sponsor representatives and Follow Through staff argued that the evaluation was too heavily oriented toward cognitive outcomes and thus would be "grossly unfair to certain sponsor approaches."[8] An Office of Education representative to the Pajaro Dunes meeting commented:

> [T]hese first meetings focused, it seemed to me, upon the objective of the follow-through evaluation and the means for accomplishing them, the notable thing being the lack of consensus on these issues at this relatively late date. As a group, the program sponsors seem not to have really accepted the idea that there will be a non-trivial evaluation in the near future. ...the hesitation of the sponsors notwithstanding, SRI seems to be moving ahead with the evaluation plans as best they can.[9]

---

[7] Krulee, pp. 247-266.

[8] Krulee, p. 219.

[9] Carl E. Wisler, Office of Program Planning and Evaluation, memorandum to Karl Hereford, Director, Program Planning and Evaluation, May 5, 1969.

In moving ahead as best they could with their ambiguous mandate, the SRI investigators seemed to try to cover all possibilities, to stretch the evaluation to cover the diverse conceptions of what Follow Through was about. Perhaps the main reason for the amorphousness of the evaluation effort was that responsibilities for shaping it were not clearly delineated. In the early years, it was assumed that the evaluation of Follow Through would be a collaborative effort involving all program participants. Apparently SRI was "expected to be responsive to the opinions of sponsors, projects, communities [and] parents as well as the Follow Through staff."[1] Indeed, Elmore characterized the collaborative problem-solving approach as the "central myth" of the early years of Follow Through.[2] As the various collaborators called for different sorts of evaluation activities, the scope of the evaluation mushroomed. SRI's proposal for the second year of evaluation (1969-70) called for additional data collection and further development of the evaluation plan, and carried a price tag of $2,150,000.[3]

Through well-recognized, ambiguity over priorities for the Follow Through evaluation continued through the year. In July 1969, controversy over the evaluation erupted during a meeting in Palo Alto. The presence of new sponsors had a "complicating effect" because they "wished to raise certain issues that had been discussed at earlier meetings.[4] In July and August, conferences to review evaluation plans turned out to be "especially stressful" as SRI came under fire for bias implicit in the measurement instruments proposed for the evaluation.

---

[1] Krulee, Organizational Analysis, p. 195.

[2] Elmore, Follow Through: Decisionmaking, pp. 231-34.

[3] Comptroller General of the United States, Report to Congress. Follow Through: Lessons Learned From Its Evaluation and Need to Improve Its Administration (Washington, D.C.: General Accounting Office, Oct. 7, 1975), p. 7.

> The evaluation plan was criticized as concentrating inordinately on classroom instructional processes and traditional measures of pupil achievement at the expense of measures, on the one hand, of affective development of children and on the other hand, of institutional change within the adult community.[5]

Ambiguity and growing discontent characterized meetings in October. As a result, SRI revamped the organization of its Follow Through evaluation effort in November. The situation was complicated by unexpectedly high costs for data collection and by preliminary results from the 1968-69 test data which, in SRI's words, "did not display evidence of strongly positive Follow Through effects." A formal review of preliminary results in February 1970, produced criticism of the evaluation's "diffuseness" and "mindless empiricism."[6]

A number of important decisions followed from the February meetings and the disappointing preliminary results. One was simply not to release the preliminary results. Another was to consider the initial year of a sponsor's involvement in Follow Through as an implementation year and hence not subject to formal evaluation. A third was to conduct the evaluation on the basis of pre- and post-testing, and so save expenses by doing away with intermediate testing.[7]

---

[4] Krulee, p. 222. SRI's work in conducting the Follow Through evaluation was complicated by the fact that they received a contract in June 1969 to conduct an evaluation of the Head Start Planned Variation Program for the Office of Child Development.

[5] SRI, Draft. Administrative History, p. F-15. See also Elmore, Follow Through: Decisionmaking, pp. 256-57.

[6] SRI, Draft. Administrative History, pp. F-18-19.

In May 1970, the Office of Education convened a panel of experts to review the Follow Through evaluation. The panel was headed by Milton Goldberg, Director of Early Childhood Education for the School District of Philadelphia. The panel recommended that SRI should continue its evaluation effort, but also advocated clarification of Follow Through program objectives and identification of policy questions to be addressed in the evaluation. The reviewers observed that:

> there seems to be some hesitancy to evaluate, in either a comparative sense--that is to compare project against project--or to evaluate against specific goals.[8]

SRI and OE bore the brunt of the reviewer's criticisms because the two institutions shared responsibilities for the bulk of the evaluation. Egbert and Snyder also, however, had funded several ancillary evaluation studies. By 1972 more than a dozen contracts and grants for such studies had been given out. Among them were awards to:

---

[7] Elmore, Follow Through: Decisionmaking, pp. 263-66.

[8] Quoted in Comptroller General, Follow Through: Lessons Learned, p. 8. For more detail on this review panel, see Elmore, Follow Through: Decisionmaking, pp. 272-77, SRI, Draft. Administrative History, pp. F-20-22; and M. Alkin, "A Review of the Evaluation of the Follow Through Program," Los Angeles, CA: California University Center for the Study of Evaluation (ED 041 642), May 15, 1970. Alkin seems to have been the most outspoken critic among the Goldberg panel members. He noted that "both the contracting office and the research institute failed to adequately specify the kind of study intended and the functions which it proposed to serve." Also, see Krulee, Organizational Analysis, pp. 271-72.

--Robert Soar, of the University of Florida, to conduct a classroom observation study in Follow Through selected sites;

--Gilbert Krulee, of Northwestern University, to study the organization of the Follow Through program and evaluation;

--Robert Fox and others at the National Training Laboratory, Arlington, Virginia, to develop a plan for facilitating local involvement in Follow Through projects;

--Joseph Grannis, of Teacher's College, Columbia University, for measurement development work;

--Thomas Sullivan of Bio-Dynamics, Inc., of Cambridge, Massachusetts, to study health service delivery in Follow Through projects;

--John Peper and Thomas McNamara, of the School District of Philadelphia, to study mobility of students into and out of the Follow Through program in Philadelphia;

--Mark Stephens of Purdue University to develop new measures of cognitive development and evaluate alternative measures of children's "locus-of-control."

A contract was even awarded to the Lockheed Missile and Space Corporation for some sort of study on early childhood programs. Most of the special studies were aimed at issues not directly assessed in the national evaluation, but it was not completely clear how the various studies might be coordinated or their findings used.[9] So the bulk of the Follow Through evaluation remained at SRI, apparently quite insulated from any of the special studies.

---

[9] Elmore observes that when the special studies reports "began to trickle in after Egbert and Snyder left, no one knew quite what to do with them. They were filed in an inconspicuous place and ignored." (Follow Through: Decisionmaking, p. 289). Reports resulting from many of these ancillary evaluation studies are described in the appendix to this report.

In June 1970, the Office of Education approved SRI's proposal for the third year of the contract, the 1970-71 fiscal year. The institute proposed that direct comparisons of the effects of different sponsors' models be postponed until each approach had been implemented for at least two years. In the meantime, SRI planned to develop information to assess the manner in which the planned variations actually varied.

To outsiders, the focus on documenting the nature of the "planned variations," after two full years of evaluation, seemed strange. Guttman and Willner later wrote:

> The spring 1970 SRI proposal lamented that the ability to develop complete and reliable descriptions of local variations represented the most important weakness to date. The programs were the central "independent variable" of the project. In layman's terms, SRI was saying that it did not know what the subject of the experiment was. SRI further explained that it had little idea of how planned variations actually varied from one another. Without adequate description of the variations, SRI explained, it would be hard to compare their effects on education. SRI promised to learn to describe the variations.[1]

The mounting problems in the Follow Through evaluation attracted attention and limited intervention from higher levels of the OE hierarchy, but SRI's proposal for the third year of evaluation was nevertheless approved, raising the total cost of the first three years of the evaluation to roughly $7 million.[2]

---

[1] Guttman and Willner, pp. 165-166. Several studies aimed at describing the planned variations had been begun prior to 1970--for example, the SRI case studies and the Soar investigation--but none had reached completion by that time.

[2] The intervention was by Karl Hereford, Director of Program, Planning and Evaluation, for the Bureau of Elementary and Secondary Education. He inserted a clause into the SRI contract designating a "procedure for transfer of data tapes in the event of contract termination" (SRI, Draft, Administrative History, p. F-23).

At about the same time as SRI's 1970-71 proposal was being approved, Follow Through Director Robert Egbert and his Chief for Evaluation, Richard Snyder, completed an annual plan for the future of the Follow Through evaluation.[3] Although the Goldberg review panel had called for a more precise specification of the purpose of the evaluation, the Egbert-Snyder plan apparently failed to further that goal. Richard Elmore described the evaluation plan thus:

> Instead of opting for a specific evaluation strategy, as the panel had suggested, the plan continued the earlier tradition of equivocation, ambiguity, and indecision. At various points it referred to the planned variation experiment as a "social engineering research strategy," a "set of field experiments," a "longitudinal evaluation," and "a series of iterations in the search for more effective ways to aid the education and general development of poor children and to achieve other social and institutional objectives of the program."[4]

The Follow Through program and its evaluation continued to be different things to different people, and indeed different things to the same people--including the Follow Through Director and Chief of Evaluation. Given the ambiguities over the main meaning of the program, it was not surprising that SRI's evaluation effort continued on an uncertain course.

By fall 1970, when SRI began to present results of pupil testing from the 1969-70 school year, the ambiguity over the purpose of the whole endeavor began to reap its harvest. In October, the SRI evaluators presented results from preliminary analyses of the 1969-70 data. The SRI Administrative History of the Follow Through Evaluation recounts:

---

[3] Cited in Elmore, Follow Through: Decisionmaking, p. 278. Memorandum from Robert Egbert to Karl Hereford, "Context and Question for Fiscal Year 1971 Evaluation Activities," June 26, 1970 (processed). Elmore comments: "Information on the file copy of this memorandum indicates it was drafted jointly by Robert Egbert and Richard Snyder."

[4] Elmore, Follow Through: Decisionmaking, p. 278.

> These data incited some controversy. ...In retrospect, no one group--USOE/FT, sponsors or SRI--can be exempted from some share of responsibility for debate that was occasionally pointless.[5]

At a November meeting in Washington, SRI presented a draft report on data from pupil testing in the 1969-70 school year and from parent interviews in spring 1970 as well. Although the two volumes of this draft report resulted from "extraordinary efforts by SRI staff...they were not considered acceptable by USOE/FT."[6] Uncertainty over the future of Follow Through was heightened when Robert Egbert announced his intention to resign as Follow Through director to take up a position as Dean of the Teacher's College, University of Nebraska, in July 1971.

A supplementary report by SRI in December 1970 was also rejected by the Office of Education. The 1969-70 report was revised and resubmitted in March 1971, but again was not accepted. Apparently, however, despite the ambiguous status of the 1969-70 draft report, a truce was reached. The draft was no more rewritten, and so not again rejected.[7]

Part of the reason for the diversion of interest from the 1969-70 evaluation results was that the Follow Through evaluation effort was coming under new scrutiny from a variety of directions, including the Government Accounting Office (GAO) and the Office of Program Planning and Evaluation (OPPE) in OE's Bureau of Elementary and Secondary Education. One incident in particular pointed up the continuing ambiguity over the nature of the program and the purpose of the evaluation.

---

[5] SRI, Draft Administrative History, p. F-25.

[6] SRI, Draft Administrative History, p. F-26.

[7] The 1969-70 report, "Longitudinal Evaluation of Selected Features of the National Follow Through Program," was never finalized, but much of the information in it was later incorporated in the 1969-71 report by Emrick, Sorenson, and Stearns.

> Mr. [Richard] Fairley, Director of the Division of Compensatory Education, visited SRI for the first time in February. When provided with an oral overview of the findings which were at that point near completion for the report, Mr. Fairley expressed dissatisfaction with the fact that pupil achievement effects could not be expressed in grade equivalent terms. It was clear from Mr. Fairley's comments that his interest was largely upon overall program effects and an identification of "best" individual sponsors to guide decisions about future Follow Through program scope. His reactions were disquieting because the report was consciously designed to minimize the chances that comparisons would be made among sponsors. In addition, since achievement measures had been constructed through the selection of elements from a variety of tests, there was no intact measure with acceptable norms from which to deduce effects in grade equivalent terms.[8]

Fairley's intervention in Follow Through reflected an ongoing and broader shift in government policies toward compensatory education. Shortly after his administration took office in 1969, President Nixon in a message to Congress had questioned the efficacy of Head Start and announced plans for transferring it to HEW. In April 1969 the Westinghouse-Ohio evaluation of Head Start was released--suggesting that disadvantaged children did not benefit greatly from attending Head Start. This generated a cauldron of controversy over the future of compensatory education.[9]

---

[8] SRI, Draft, Administrative History, p. F-28.

[9] New York Times, February 20, 1969. For the fuller story of the Westinghouse-Ohio report and the controversy surrounding it, several accounts are available. For the con view see Smith and Bissell, "Report Analysis," Harvard Educational Review, 40:1(1970), 51-104. For the proposition see V. G. Cicirelli, J. W. Evans and J. Schiller, "A Reply to the Report Analysis" in the same issue, pp. 105-129. For a more middling account see S. White, "The National Impact Study of Head Start" in Disadvantaged Child Vol. III, Compensatory Education: A National Debate, J. Hellmuth, Ed., (New York: Bunner-Mazel, 1970), pp. 163-184. A more recent account is Lois-ellin Datta, "The Impact of the Westinghouse/Ohio Evaluation on the Development of Head Start: An Examination of the Immediate and Longer-Term Effects and How They Came About," in Clark Abt (ed.), The Evaluation of Social Programs (Beverly Hills, CA: Sage, 1976), pp. 129-91.

The controversy over compensatory education was rekindled in March 1970, when in a special message to Congress Nixon observed that compensatory education, including Title I of the Elementary and Secondary Education Act of 1965, had not measurably helped disadvantaged students and should not be expanded, pending further study.[1] In October the Administration proposed a cut in Head Start funds and amidst continuing controversy a "Head Start Crisis Day" was declared and nationwide lobbying efforts begun to boost support for compensatory education.[2]

In the course of such controversies, the Office of Education seemed to be steadily shifting more toward the view of Follow Through as a planned variation experiment and away from the interpretation that it was a trial phase for a nationwide service program. In his "Follow Through" paper apparently written in early 1971, Robert Egbert had suggested that a "large-scale Follow Through type program" would commence in 1975-76.[3] Yet later in 1971, HEW Assistant Secretary for Planning and Evaluation Laurence Lynn wrote to HEW Secretary Elliot Richardson, "Follow Through is either an experimental or service program, depending on who you talk to. We believe it has been, or should have been primarily the former."[4]

---

[1] New York Times, March 3, 1970.

[2] New York Times, October 30, 1970, December 13, 1970.

[3] Egbert, "Follow Through," ca. 1971, p. 23.

[4] Laurence Lynn memorandum, "Weekly Report" to Elliot Richardson, July 17, 1971 quoted in Elmore, Decisionmaking, p. 291.

As the main meaning of Follow Through in the minds of HEW officials seemed to shift, dissatisfaction over the evaluation grew apace. In January 1971 the Office of Education instituted an independent review of the SRI evaluation effort, by the Pacific Educational Evaluation Systems (PEES). The PEES review criticized the evaluation on a variety of grounds, including:

--the SRI project was understaffed at upper levels;

--the evaluation project had grown so large as to be perhaps unmanageable;

--the effort lacked clear definition of expected reports and a "schedule for the major steps in their production";

--more generally, the evaluation had been troubled by a "collaborative" or "joint-venture" method of decision-making.[5]

The GAO review of the Follow Through evaluation effort reached similar conclusions. Though not officially published until August 1971, the GAO report was available in draft form within the Office of Education, and evidently widely circulated, as early as May. The GAO concluded, "The $7 million study of the national Follow Through program is another example of the need for more specific statements of contract objectives and requirements."[6]

---

[5] These criticisms from the PEES review are summarized from Krulee's Organizational Analysis of Follow Through, pp. 280-283.

The growing criticism of the Follow Through evaluation raised doubts as to whether the SRI evaluation contract would be extended into the 1971-72 fiscal year. Indeed, one of the goals of the PEES review was to prepare a request for competitive bids from other research firms to conduct the evaluation. The PEES reviewers concluded, however, that the contract with SRI ought to be continued because: 1) it did not seem prudent to change contractors "at the half way point in the evaluation study" and 2) "there are few contractors in this country with the capability and knowledge to conduct the Follow Through evaluation."[7]

In effect, SRI was given a reprieve, another chance to come up with an evaluation report that would be acceptable to the Office of Education. So, despite the continuing ambiguity over exactly how the evaluation should be conducted, the multi-million dollar SRI proposal for 1971-72 was approved on June 29, 1971. An important proviso in this approval was that SRI's contract would be extended only tentatively pending further negotiation and specification of work to be performed under the contract. SRI was to submit documentation to clarify its proposal by the middle of September.

This interval, the summer of 1971, was a period of crisis and growing ambiguity for the FT program as well as the evaluation. Robert Egbert left his position as Follow Through director at the end of June 1971, and at the same time Don Rose of BESE's Office of Programming Planning and Evaluation (OPPE) was assigned responsibility for reviewing the SRI contract. OPPE had intervened directly in the negotiations for the new SRI contract and Rose's designation as project monitor was clear indication of OE's intent to straighten out the evaluation effort. Shortly after these changes, Rosemary Wilson was named to replace Egbert, and in August 1971, Richard Snyder stepped down from his position as chief of the Follow Through evaluation.[8]

---

[6] Comptroller General of the United States, Report to Congress: Need for Improving the Administration of Study and Evaluation Contracts (Washington, D.C.: GAO, August, 1971), p. 16.

[7] Quoted in Krulee, Organizational Analysis, p. 283.

On top of this flux in leadership, uncertainty over the future of Follow Through was heightened even further at a July meeting hosted by SRI. The meeting was attended by Follow Through and other OE officials, by sponsor and project represenatives, by state education officials and by Follow Through parents. At the gathering, SRI presented plans for the coming year of evaluation and preliminary findings of previous data analysis. Attention was diverted from SRI's evaluation efforts, however, by discussion of Richard Fairley's long-range plan for Follow Through. Fairley's plan called for "phased introduction of State-administered Follow Through/ Title I match grant programs in all states that will extend Follow Through-tested successful practices" to the Title I program of the Elementary and Secondary School Act of 1965.[9] The idea was enticing in a couple of ways. The notion of introducing "Follow Through-tested practices" into the Title I program promised an air of legitimacy to an effort which had recently been challenged as ineffective and even as flagrantly violating the law.[1] At the same time, Fairley's idea provided a raison d'etre for Follow Through--the results of the planned variation experiment were to be used to improve the Title I program.

---

[8] SRI, Draft Administrative History, pp. F-28-29; Krulee, Organizational Analysis, pp. 284-85.

[9] Krulee, Organizational Analysis, p. 44.

Fairley had also supported the SRI production of a "Consumer's Guide" to Follow Through.[2] The Guide was to be prepared in draft form by late 1971 and was scheduled for publication in the spring of the next year. The idea was that the Guide would provide information on the alternative educational approaches developed in Follow Through for potential consumers of the educational models. The Consumer's Guide was to make use of information developed in the national evaluation, but would be written in non-technical language so as to help parents and other citizens choose educational approaches to further the objectives they sought.

There were two main difficulties with these ideas. First, the Follow Through evaluation had produced no definitive information on the effectiveness of alternative models of the sort Fairley envisioned for his Consumer's Guide. Second, Fairley's proposals were based on the view that Follow Through was primarily a planned variation experiment designed to test alternative models of education along common dimensions. Since not all stakeholders in the Follow Through program shared this view, it was natural that Fairley's views provoked consternation. As Gilbert Krulee observed, "To the extent that [Fairley's plan] was perceived as a redirection of Follow Through, sponsors and local project representatives reacted with expressions of annoyance and anxiety."[3]

---

[1] J. Murphy, "Title I of SEA: The Politics of Implementing Federal Education Reform," Harvard Educational Review, 41 (February, 1971), pp. 35-63.

[2] The SRI Administrative History (p. F-31) notes that the idea of a consumer's guide to Follow Through evolved over a matter of time. Apparently it came to attention in July 1971 when Fairley proposed his five-year plan because it meshed with his ideas for expanding Follow Through. See Krulee, Organizational Analysis, p. 288.

At a meeting in Ypsilanti, Michigan in October, 1971, Follow Through sponsors made their view explicit:

> We hereby propose and urgently recommend that the Office of Education commission a task force to undertake the study of Follow Through as a national experiment in developing educational alternatives.[4]

The sponsors proposed that the report of this task force "serve as a companion volume to the proposed Follow Through Consumer's Report" and that it be "as widely distributed as the latter."[5]

From SRI's point of view, the evaluation effort in the summer of 1971 seemed to be degenerating into a quagmire. The SRI Administrative History recounts that, by late August, almost "all USOE/FT staff with whom we had worked closely during prior years had resigned or been reassigned." Also, with the shift of contract administration to the OPPE of BESE, and in contrast to the three preceeding years, "relationships with USOE/FT became highly formalized."[6]

---

[3] Krulee, Organizational Analysis, p. 287.

[4] High/Scope Educational Research Foundation, Working Conference on Evaluation of Follow Through Objectives Not Measured by Academic Tests, Report of Conference Activities and Recommendation (Ypsilanti, MI: High/Scope, October, 1971), p. 8.

[5] High/Scope, Working Conference, 1971, p. 9.

[6] SRI, Draft, Administrative History, p. F-30. Ray Rhine, one of the SRI evaluators, quotes an evaluation project leader at the time confessing that he would welcome a "sub-optimum escape route from this miserable morass" ("Toward More Effective Utilization of Behavioral and Social Scientists in School-Community Intervention Programs," paper presented to the American Education Research Association Meeting, Chicago, April 17, 1974.)

Despite the growing tension over the course of the evaluation and the main purpose of the program, SRI continued during the summer months of 1971 to prepare for data collection in the fall, to work on specification of future evaluation activities, and to prepare "a cascade of successive substantive reports of findings." [7]

In mid-September, SRI delivered to USOE two documents--one elaborating data collection plans for 1971-72 and the other outlining "analyses thought to be appropriate to the objectives of preparing the Follow Through Consumer's Guide." Reviews of these documents by OE officials and outside consultants were highly critical, and the reports were deemed unacceptable. A consolidated critique of the September reports ran to seventeen single-spaced pages.[9] Reviewers had dozens of concerns: regarding the design, sample and analyses for the planned variation experiment; the plans for pupil testing, classroom observation and parents interviews, as well as program description, cost analyses, reporting, budgeting and staffing. Altogether they made 45 separate recommendations for improving the evaluation.

The dissatisfaction over the September reports led to a series of meetings between SRI and OE representatives in October and November and to negotiations over precisely what kind of reports SRI would prepare in the future. However, before SRI could go much further in preparing additional reports, events intervened which sharply altered the course of the Follow Through evaluation.

---

[7] SRI, Draft Administrative History, p. F-31.

[8] SRI, Draft Administrative History, p. F-32.

[9] J. Crooks and D. Burns, "Consolidated Review of September 17 Deliverables," memorandum to F. Bresnick, Acting Chief, Research and Evaluation Section, Follow Through Program, October 6, 1971.

Revamping the Evaluation and the Shift to Abt

In November, 1971, Garry McDaniels joined the Office of Education staff as a replacement for Richard Snyder. McDaniels came to the Office of Education from the Institute for Child Study at the University of Maryland, where he had been an Associate Professor. Although new to government service, he quickly took the lead in revamping the Follow Through evaluation. His review of the SRI evaluation focused on several issues:

--the joint venture or collaborative mode of decision-making between SRI and OE was not working to produce an effective Follow Through evaluation design and analysis.

--SRI's data collection activities had been generally successful and ought to be continued.

--SRI's capabilities in the realm of instrument development were "quite limited" and work in this realm ought to be curtailed. Also, the review concluded that USOE ought to assume responsibility for selection of instruments to be used in the Follow Through evaluation.

--SRI's performance in analyzing Follow Through evaluation data had been inadequate and a decision regarding the analysis of the 1971-72 data should be deferred pending receipt of the February 1973 report.[1]

McDaniels' critique reflected concerns over the evaluation which had been voiced for several years; in the February 1970 review, by the Goldberg review panel, by GAO investigators and most recently by the reviewers of the September SRI reports. Over the next several months McDaniels transformed these concerns into several basic changes in the national evaluation.

A difficulty that underlay many of the other problems in the Follow Through program and in the national evaluation was staffing. Robert Egbert had identified staff support as "one of the most persistent problems of Follow Through" and in congressional hearings in February 1971, the Office of Education had requested funds for additional staff for the program end of Follow Through operations.[2] Since the Office of Education was eliminating the "joint venture" mode of decisionmaking with SRI and was taking on more evaluation planning in-house, the staffing problem promised to get even worse.

---

[1] Follow Through, Internal Memorandum of December 10, 1971, cited in Krulee, Organizational Analysis, pp. 292-293.

McDaniels remedied the situation by awarding a contract to the Huron Institute of Cambridge, Massachusetts in December 1971.[3] The key figures in early Huron work on Follow Through were David Cohen and Marshall Smith, both then faculty members at the Harvard Graduate School of Education. As one of the participants in the 1967 White House Task Force on Child Development, Cohen had advocated the idea of planned variations in education, and Smith was then project director for Huron's analyses of Head Start Planned Variation data. In a memo to McDaniels in late 1971, Smith outlined a "seductively simple" strategy for salvaging the evaluation:

> 1. Obtain within OE (and HEW) as close to a consensus as possible about the long-term objectives of the Follow Through evaluation. These objectives must be defined to a degree where they can be translated into design specifications. The objectives, rather than shortterm political and financial considerations, must determine changes in the design...
>
> 2. The second part of the plan would be to formulate a design (within the existing financial and programmatic constraints) to maximally meet the specified objectives....

---

[2] Egbert, "Follow Through" ca. 1971, p. 60 and U.S. House of Representatives, Hearings Before the Committee on Appropriations, February 18, 1971, p. 284.

[3] Huron's contract was awarded on a sole-source basis, that is, it was not bid competitively. This was deemed justified because of Huron's experience in analyzing the evaluation data from the Head Start Planned Variation program.

> 3. The third part of the plan is to insure that the data collected be as reliable as possible...
>
> I believe that a successful implementation of this plan would save the Follow Through evaluation.[4]

One of the most pressing problems for the evaluation was that there had been no overall rationale for previous evaluation testing. Instead, the evaluation sample had simply accrued out of a variety of different kinds of decisions--variously having to do with administrative convenience, sponsor preferences, instrument development considerations, and cost constraints. Therefore, one of the primary purposes of the Huron contract was to design a systematic plan for sample selection.

The selection of sites for inclusion in Follow Through evaluation testing was severely constrained by the previous testing. By January 1972, when the Huron sample design was formulated, four cohorts of children had already entered the Follow Through planned variation programs.[5]

---

[4] Marshall Smith, memorandum to Garry McDaniels, processed, November 10, 1971, pp. 1-3.

[5] Recall that a decision was reached to consider the original group or or cohort of children entering the Follow Through planned variation program in 1968-69 as an implementation cohort and not subject to evaluation. Thus the group entering in 1969-70 was labeled cohort I, that entering in 1970-71 cohort II, etc. For a diagram showing the relations of the Follow Through cohorts, see Figure 3.

Within the constraints of previous testing, a sampling plan was developed by Jane David of Huron, who had studied with Smith at Harvard. David's plan was guided by two objectives:

1. The first objective was to maximize the potential for valid comparisons both among Follow Through sponsors and between Follow Through and comparison classes.
2. The second objective was to facilitate the determination of how much, when and for whom compensatory education such as that delivered by Follow Through is effective.[6]

Operationally, this led to an attempt to include in future testing "at least two similar sites per sponsor in cohorts I and II and at least five sites per sponsor in cohort III."

The clear implication of these decisions was that, from that point on, the Follow Through evaluation would be guided by the planned variation vision of Follow Through. Though McDaniels commissioned a number of auxiliary research studies,[7] the thrust of the Follow Through national evaluation under his leadership quickly became evident--a comparison of the effects of alternative curriculum models on children, with comparisons made both between models and between the Follow Through sponsored models and regular, that is non-Follow Through, elementary school curricula.

---

[6] Jane David, "Rationale for Selection of Sites and Grade Levels for Spring 1972 Follow Through Testing," Huron Institute, January 20, 1972. David's sampling plan was also guided by a number of secondary considerations: (1) Sites having Head Start Planned Variation Projects were to be included in the sample for Spring, 1973, according to an agreement between OE and OCD; (2) Sites needed for other special studies, for example, the SRI classroom observation studies and the Huron summer study, were also to be included.

[7] In addition to its technical assistance, McDaniels contracted Huron to conduct studies on the impact of summer experience on students' academic growth, on units-of-analysis issues, on the pros and cons of longitudinal versus cross-sectional evaluation studies, and on a number of other topics. Reports resulting from these studies are described in the appendix.

In pushing the evaluation toward the planned variation conception of Follow Through, McDaniels had clear support from his superiors in HEW. During the controversy over Follow Through in the summer of 1971, after concluding that Follow Through should have been an experimental program, HEW Assistant Secretary Lynn had prescribed that "Both the design and evaluation [of Follow Through] need to be improved considerably for it to be a decent experimental program. We will be pursuing this objective over the next two months."[8]

After reviewing SRI draft reports in January and February 1972, OE officials decided that the next contract for analysis of Follow Through evaluation data should not necessarily go to SRI but instead should be opened for competitive bidding. The Request for Proposals (RFP) for future analysis was announced in March 1972, and clearly reflected the sharpened focus for the evaluation effort. The principal issues to be addressed in the Follow Through evaluation according to the 1972 RFP were:

--Is there any overall difference in the impact of Follow Through and comparison schools? That is, *does the Follow Through Program in the aggregate have a greater impact on parents and children than non-Follow Through schooling*?

--Are there *educationally significant differences* in the impact of the various educational strategies being tested in Follow Through?

---

[8] Laurence Lynn memorandum "Weekly Report" to Elliot Richardson, July 17, 1971, quoted in Elmore, *Follow Through: Decisionmaking*, p. 291.

--If such differences exist among the treatments, are they consistent over time--that is, do some treatments consistently show better results than others, or do the differences shift from year to year?

--For any given treatment, do differences in the completeness of its implementation have any relationship to differences in impact?

--Are differences in impact related to the characteristics of children, such as age, socio-economic status, pre-school experience?

--Are there interactions between the characteristics of children and the characteristics treatments which affect impact?

--Are there interactions between child characteristics and program characteristics which are not model specific (e.g., exposure time)?[1]

The RFP noted that these questions--related to the performance of children and to the attitudes and behavior of parents, teachers, and aides--comprised only a subset of the "broad goals of Follow Through."[2] In other words, in refocusing and redesigning the evaluation, McDaniels was narrowing in on a particular subset of the questions which had been asked of the Follow Through evaluation. Gone from the new focus were inquiries like SRI's community studies and the efforts to develop new instruments.

Thirty-one firms responded to the RFP, and seven, including SRI, actually submitted proposals.[3] The winner of the competition, selected in June, was Abt Associates, Inc., of Cambridge, Massachusetts. The Abt proposal dealt mainly with the problems which the analysis would have to overcome such as:

---

[1] OE, Request for Proposal, Follow Through, Draft, March 13, 1972. This draft is virtually the same as the final RFP issued.

[2] OE, RFP, Follow Through, Draft, March 13, 1972, p. 8.

[3] HEW Audit Agency, Draft Report on Review of Administration of and Performance Under Contract with the Stanford Research Institute, Office of Education and Office of Child Development, December, 1972, p. 27.

the development of appropriate measures, the problems of missing data, ascertaining the accuracy of data, the generalizability of results, and the issue of middle-class bias in interpreting data. The Abt planners noted:

> Given the character of the Follow Through data base, it is much easier to catalogue the constraints and hinderances that must afflict generalizations from it than to specify precisely what kinds of generalizations will be possible and trustworthy.[4]

But despite the "awesome range of problems faced in dealing with the evaluation of a large-scale multidimensional educational program," Abt expressed optimism regarding the "task of evaluating the Follow Through program." According to Abt there were two main reasons for optimism. First was the Follow Through program's "careful attention to design problems...both before and during the operation of the program." The other cause for Abt's positive outlook was the view that, "over the past five years or so there has been a proliferation of statistical applications using techniques which allow one to deal with complex data collected under quasi-experimental conditions... in particular...the increasing use of the general linear model of multivariate statistics."[5] Despite its lengthy treatment of the problems inherent in the analysis of the Follow Through evaluation data, the Abt proposal was highly curious in one respect. Not until more than halfway through the technical proposal was there any mention of the overall purpose of the evaluation. Buried in a discussion of possible middle-class bias in the evaluation, the proposal noted:

---

[4] Abt Associates, "A Proposal to Conduct an Analytic Study of the Follow Through Program, 1969-72. Part I: Technical Proposal." RFP No. 72-37, AAI #1507, May 5, 1972, p. 19.

[5] Abt, "Proposal", May 5, 1972, p. 1.

> The basic purpose of the Follow Through Study, and hence all the analyses, is limited to determining which of a number of different approaches is most effective, under what conditions, and related to (which) individual characteristics of children within the sample population.[6]

The 1972 revamping of the evaluation also encompassed the troublesome problem of instrumentation--what test instruments should be used to assess effects of Follow Through on participating students. The instrumentation issue had a long history in Follow Through. How to choose a set of test instruments which would provide a fair assessment of diverse sponsors had bothered Follow Through evaluation planners from the start. Even Kurfeerst and Stephens had bemoaned the limitations of standardized testing and the "difficulty of obtaining a truly culture-free test."[7] The response to this problem during the SRI era of the Follow Through evaluation had been to try to develop new test instruments. A special section of the SRI project had been devoted to the development of noncognitive measures, and in the early years SRI administered a wide range of achievement test items suggested by sponsors. The result of this nonstandardized testing was that in the early years of the evaluation, there were few common measures which could be used to compare outcomes across all sites or across different waves or cohorts of Follow Through students.[8]

McDaniels' strategy was to cut his losses, so to speak, and, in order to give the evaluation some chance of success, to make do with tests that were already developed and that seemed to have the fewest shortcomings. Toward this end, a panel of experts was convened in early 1972 to select tests for use in spring 1972 and during the following school year.[9] The battery selected for testing Follow Through children at the end of the program had two parts:

---

[6] Abt, "Proposal," May 5, 1972, p. 42.

[7] Kurfeerst and Stephens, p. 5.

[8] The lack of standardized testing was a major problem in early Follow Through evaluation studies. See, for example, Robert Soar, Final Report Follow Through Classroom Process Measurement and Pupil Growth (1970-71) (Gainesville, FLA: University of Florida, June, 1973.)

Cognitive Domain--Metropolitan Achievement Test (MAT);
Ravens Progressive Matrices.

Affective Domain--Coopersmith Self-Esteem Inventory (Revised);
Intellectual Achievement Responsibility Scale (Revised).

At intermediate points in the program, students were to be tested with only the MAT.

So, by the end of June 1972, the direction of the Follow Through evaluation had shifted enormously and focused more narrowly on the planned variation aspects of the program. With the help of Jane David of the Huron Institute, McDaniels had redesigned the sampling plan for testing Follow Through children, and with the aid of a special review panel, he selected a set of standard test instruments on which to base the evaluation. From then on SRI's main task would be to collect data and to analyze and report on only the classroom observation data. Abt would have responsibility for analyzing student test data and all information from teacher and parent questionnaires. An HEW audit of the evaluation summed up the shift in responsibilities as follows:[1]

---

[9] The panel consisted of: Dr. Lois-ellin Datta of OCD, Dr. David Weikart of High/Scope Foundation (A Follow Through curriculum model sponsor), Dr. Vicki Shipman of the Educational Testing Service, and Dr. Ed Epps of the University of Chicago.

[1] HEW Audit Agency, Draft Report, December 1972, p. 27. The Audit Agency summary was slightly incomplete in two respects. SRI continued analysis activities beyond mid-1972 in the form of Jane Stallings' classroom observation study, and Abt participated in redesigning the parent interview and teacher questionnaire forms after they received the 1972 contract.

Task	Past Responsibility	Current Responsibility
Instrument Construction	SRI/OE	OE
Battery Selection	SRI/OE	OE
Analysis	SRI	Abt
Sample Selection	SRI/OE	Huron
Data Colletion	SRI	SRI

The list shows how much responsibilities for the Follow Through evaluation had shifted in 1972: OE's role was more prominent; two new organizations--Abt and Huron--now played parts in the evaluation effort; and SRI's role was sharply reduced. From 1972 onward, SRI's main responsibility in the evaluation effort would be data collection--a function which reviewers generally agreed SRI had carried out successfully in the past. After SRI responsibilities were cut back, Richard Marciano who had been closely associated with SRI's successful data collection and processing moved up in the fall of 1972 to become director of the SRI Follow Through project. The list of shifts in responsibilities is highly curious in one respect. It encompasses all of the duties included in the original 1968 SRI contract save one--the responsibility for "working with sponsors to specify the content and objectives of their models."

In the midst of McDaniels' reorganization, the Follow Through evaluation caught the attention of two investigators from Ralph Nader's Center for the Study of Responsive Law--Daniel Guttman and Barry Willner. On May 24, they wrote to HEW Secretary Elliot Richardson charging OE with waste and abuse of regulations in the SRI-FT contract. Even though SRI reports had been deemed "valueless," Guttman and Willner charged, the SRI contract had been extended by OE Follow Through personnel "because (1) they want to save face and (2) they want to avoid 'losing' the 6.2 million dollar balance of contract funds by having it revert to the treasury."[2]

Reported in Education Daily, the Guttman-Willner charges brought the Follow Through evaluation under new scrutiny. In response to Guttman and Willner's critique, Richardson commissioned three reviews of Follow Through: by the Follow Through branch, by the HEW Office of Planning and Evaluation, and by the HEW Audit Agency. McDaniel's reply to the Guttman-Willner letter noted that most of what had drawn the investigators' wrath had already been remedied, and that most of the important remaining problems in the evaluation were not even addressed by Guttman and Willner.[3]

The HEW Audit Agency response took considerably longer than that of the Follow Through branch. A December, 1972 draft of its report stated

> Because the services required of SRI were of a highly complex and technical nature, the Audit Agency was not qualified to evaluate independently the quality of SRI's performance. Consequently we relied on the results of reviews made by various panels, composed of experts from within and outside the Government, convened to assess SRI's performance under the contracts. The general consensus of the review groups what that SRI's efforts on data collection were performed satisfactorily. In the areas of data design, test development and data analysis and reporting SRI's performance was not considered entirely satisfactory because of limited staff capabilities.[4]

---

[2] Daniel Guttman and Barry Willner, Center for the Study of Responsive Law, Washington, D.C., letter to Elliot Richardson, Secretary of HEW, May 24, 1972. Guttman and Willner's critique of SRI's evaluation contract was repeated, in large part, in their book The Shadow Government, (NY: Pantheon Books, 1976). The allegations by Guttman and Willner concerned the SRI contract with OCD for the evaluation of Head Start Planned Variation as well as that of Follow Through.

[3] R. Elmore, FT: Decisionmaking, p. 310.

[4] HEW Audit Agency, Draft Report, December, 1972, p. 3.

SRI was not, however, the only target of the auditors' critique. The audit agency concluded,

> --OE did not give prospective contractors sufficient time, considering the magnitude and complexity of the Follow Through project, to submit proposals. Less than two months elapsed between the date that preliminary proposals were requested and the contract was awarded.
>
> --The basis for the selection of SRI was not documented in OE files as required by Federal Procurement Regulations.
>
> --The initial OE and OCD contracts did not contain specific statements of contract objectives and requirements.[5]

Of the three reviews of Follow Through resulting from the Guttman-Willner allegations, the third one looked most closely at the substantive problems facing the evaluation. This inquiry was conducted by a panel of reviewers--including HEW officials, outside experts, and Follow Through parent representatives. The review panel met four times during the fall of 1972, and issued a final report in March 1973.[6]

Lamenting the complexity of the issues surrounding Follow Through the panel disclaimed "...it is hard to make recommendations which are not gross oversimplifications." Nevertheless, the panel set out two guidelines as central to all its proposed recommendations:

---

[5] HEW Audit Agency, Draft Report, p. 2.

[6] Office of Education, Second Report of the Follow Through Evaluation Review Panel (processed), March, 1973, 56 pp. "(A) 'technical core' group, which (is) responsible for synthesizing all inputs into the panel into a final set of recommendations, included Dr. Milton Goldberg, Executive Director of Early Childhood Programs in Philadelphia; Dr. Lilian Katz, Professor at the University of Illinois and Director of the Eric Clearinghouse on Early Childhood; Dr. Edgar Epps, Professor at the University of Chicago; and Ms. Deborah Klein Walker, Special Education Consultant in the Office of the Assistant Secretary for Planning and Evaluation. The resource members of the panel are Ms. Helen Testamark Ms. Doris Cohn, and Mr. Thomas Murphy, Follow Through parent representatives; Dr. Nick Rayder, sponsor representative; Mr. Eugene Tucker, Office (continued on the next page)

(1) Whenever possible, evaluation efforts should take into account all components of the Follow Through program. This can best be done only for instructional approaches to early childhood education since these were the only elements systematically varied in the design of the national experiment. In pursuit of successful strategies for compensatory education of disadvantaged children, the major question to be addressed in the Follow Through evaluation efforts is: What works for whom under what circumstances: In other words, what part(s) of the Follow Through Program are successful for which children in which places at what grade levels?

(2) There is a legitimate and meaningful role for various groups (i.e., national government, local organization, sponsors, and parents) in the evaluation of Follow Through. All of the evaluation efforts should be integrated into a synthetic whole which produces information for making both formative and summative educational decisions. This integration effort should be accomplished by the Follow Through Evaluation Branch.[7]

Apart from these general guidelines, the review panel adhered quite closely to the McDaniels-Huron plan for the evaluation. The panel report described Jane David's work in designing the evaluation testing sample and cited the same "two main objectives of the national Follow Through evaluation" which had guided her work. Regarding the evaluation design, the review panel concluded, "The basic evaluation strategy implemented by the Follow Through Evaluation Branch is sound."[8]

---

[6] (Continued from preceding page) of Program Planning and Evaluation Office of Education; Dr. David Iwamoto, Bureau of Elementary Education and Secondary Education Office of Education; and Dr. Garry McDaniels, Acting Chief, Follow Through Evaluation, Office of Education."

[7] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, p. 3.

[8] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, p. 18.

The panel's recommendations regarding local and sponsor evaluations were more ambiguous. For local communities the panel report merely listed a dozen areas of inquiry that local community evaluation might pursue--though it noted that "many communities will have neither the capacity nor the resources to accomplish all of these activities."[9] When it came to evaluation by Follow Through sponsors, the ambiguity became pure murk. This section of the report began with a statement of the "wide range of sponsor positions on various evaluation issues." For instance,

> Some sponsors feel a test battery can be developed to assess across sponsor differences; other sponsors feel the whole concept of testing is foreign to their programs.

In nineteen areas, the report illustrated the diversity of opinion with the "Some sponsors believe A, while others believe B" format. On only two evaluation issues was there apparent agreement among the sponsors:

> --All sponsors feel evaluation should be conducted by all stakeholder groups; [and]
>
> --Follow Through sponsors feel evaluation is an integral part of program development and operation.[1]

After remarking on the need for sponsor specification of intended outcomes, target populations, and program attributes and conditions, the panel recommended:

---

[9] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, p. 18.

[1] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, p. 18.

> . . . the continuation of the priorities for sponsor's evaluations that were established during the past year by the Follow Through evaluation staff. They are as follows:
>
> 1. Small sponsors who receive relatively limited examination in the National Evaluation should include a broad range of documentation. Data on achievement should be a major concern in these efforts.
>
> 2. For all sponsors, proposals which deal primarily with the gathering and analysis of existing data are encouraged.
>
> 3. Sponsors should order their objectives. This ordered list should be compared to the objectives established by the test battery in the National Evaluation. Where crucial sponsor objectives are not sampled by the National Evaluation, sponsors should propose a documentation effort.
>
> 4. Lowest on the list of activities for sponsors to pursue is instrument development. There are two reasons:
>
>   a. Past experience with sponsors, SRI and other contractors indicates that this is a difficult, costly, and time-consuming task with limited probability of success.
>
>   b. Since the time remaining for one to document the impacts of the program is extremely time limited, there should be concentrated efforts on more immediately achievable goals.[2]

In addition to its review of national, sponsor, and local evaluation plans, the panel also briefly recapped issues concerning instrumentation and the implementation of sponsors' curriculum models.

---

[2] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, pp. 19-20.

The only other section of the panel's report dealt with parents' role in Follow Through evaluations. A scant eighty-six words long, this section expressed the panel's strong recommendation "that parents be a part of all Follow Through evaluation efforts," and promised that "the panel is presently reviewing a variety of specific functions the parents can play in all aspects of the Follow Through evaluation."[3]

The brevity and wording of this recommendation is highly curious since it came only five months after a major confrontation over the role of parents in the Follow Through program and its evaluation. This confrontation, remarkably reminiscent of the Atlanta conference of 1968, occurred during a national Follow Through conference in Denver in October 1972. The week-long conference was attended by 700 to 1100 people, including Office of Education and sponsor representatives, local project directors, and parents of Follow Through children. An Education Development Center staff member described the highlights of the conference:

> Tuesday evening things began to get hectic--there was a sponsors' meeting in one area and a parents' meeting in another...
>
> First the sponsors decided to band together in full support of the parents. A written resolution was drawn up in the wee hours of the morning for all the sponsors to sign and to be read to the full convention at the appropriate time.
>
> The parents' meeting was indescribable--the efforts, motivation, involvement and true concern, defy words. By dawn, a statement had been drawn up demanding the

---

[3] OE, Second Report of the Follow Through Evaluation Review Panel, March, 1973, p. 21.

> continuation of Follow Through, again to read to the full convention at an undetermined time....
>
> Things reached a climax on Thursday morning. Ms. Helen Testamark, chairman of the Steering Committee for the PAC national group, read the FT parents' statement... to the full gathering of convention participants, followed by the sponsors' statement read by David Weikart (Chairman of the Sponsors' Group). Both statements were endorsed as resolutions by the entire conference.[4]

As a "first order of business," the parent resolution declared the Denver Conference inconsistent with Follow Through guidelines because parents of Follow Through children had been excluded from full involvement in its planning. The parents also expressed concern for the continuation of Follow Through and parent participation in it. On the topic of evaluation, the parent resolution was especially forceful:

> DESPITE THE AMOUNT OF MONEY THAT HAS BEEN GIVEN TO STANFORD RESEARCH INSTITUTE TO PRODUCE THE DATA TO DOCUMENT THE WORTH OF THE FOLLOW THROUGH PROGRAM, SRI HAS YET TO PROVE BEYOND A SHADOW OF A DOUBT THAT OUR CHILDREN ARE LEARNING. WE KNOW THAT OUR CHILDREN ARE LEARNING. WE KNOW THAT THEY CAN LEARN. WE HAVE THE EVIDENCE THAT THAT ARE LEARNING AND CAN LEARN. NO ONE HAS EVER ASKED US TO DEMONSTRATE THAT FOLLOW THROUGH IS WORKING. WE RECOMMEND/DEMAND THAT THE CONTRACT WITH SRI BE TERMINATED IMMEDIATELY. WE RECOMMEND/DEMAND THAT A MECHANISM BE DEVELOPED IN WHICH THE EVALUATION PROCESS IS MANAGED AND CONTROLLED BY THE PARENTS THEMSELVES. PARENTS MUST BE INVOLVED IN THE COLLECTION AND INTERPRETATION OF DATA ABOUT THEIR OWN CHILDREN--AND BE EMPLOYED AS SELF-EVALUATORS THEMSELVES....WE WILL DESTROY

---

[4] Jeanne McDonald, "Denver National Follow Through Conference," Voices (newsletter of the EDC Follow Through Program), December, 1972, P. 10. "PAC" is an acronym alternatively deciphered as the Parent Advisory Committee or the Policy Advisory Committee. By whichever name, each Follow Through site was supposed to have one of these committees in order to provide a forum for parent participation in local Follow Through program decision-making. At the time of the Denver conference there was a move afoot to organize a national committee to represent Follow Through parents.

> ANY PROGRAM THAT ATTEMPTS TO EXPERIMENT UPON OUR CHILDREN BASED UPON THE DEFINITIONS OF OTHERS. WE WILL INSIST UPON CERTAIN PRACTICES THAT DO NOT FIT WELL WITH STRICT RESEARCH IDEAS. FOR INSTANCE, WE WILL OPPOSE PROGRAMS FOR PEOPLE IN NEED THAT SELECT ONLY A FEW TO GET IMPROVED SERVICES, OR THAT LIMIT THE HELP GIVEN TO ONLY THAT WHICH IS MEASUREABLE. THE BURDEN OF BEING ABLE TO MEASURE THE KIND OF PROGRAM WE WANT SHOULD BE THAT OF THE EXPERTS--RATHER THAN TO DESIGN A PROGRAM THAT THEY KNOW HOW TO MEASURE.[5]

The Follow Through sponsors concurred. They proclaimed "One truth emerging from our experience in Follow Through is that the involvement of parents in decisions about their children is essential for effective education." The sponsor resolution presented to the conference made it clear that those who viewed Follow Through as more than a planned variation experiment would not rely solely on the evaluation to further their cause:

> . . .be it resolved that Follow Through Sponsors, PACs [Policy Advisory Committees] and other Stakeholder groups and friends of Follow Through join together to take the necessary action to present the case to those responsible for legislation affecting young children so that the concepts of Follow Through be legislatively sanctioned in the next session of Congress as a basis for the maintenance and extension of these benefits of Follow Through to children who need them.[6]

What had triggered the parents' and sponsors' resolve was a message from the Office of Education that Follow Through, as an experimental program, would "be brought to an orderly conclusion."[7] By focusing on a single conception of Follow Through--Follow Through as an experiment --and drawing a natural conclusion from that view, the OE communication directly threatened "stakeholders" in Follow Through who held other conceptions.

---

[5] Voices, December, 1972, pp. 12-14.

[6] "Sponsor Resolution, Denver National Follow Through Conference," quoted in Voices, December, 1972, p. 16. Policy Advisory Committees or PAC's were mandated for every FT project in the original FT program manual. At least 50% of each PAC were to be parents of FT children and the PAC was supposed to have a "substantial role in the planning and management of the Follow Through project." OE, Draft Follow Through Program Manual, February 24, 1969, p. 6.

The controversy at the Denver conference was highly reminiscent of the flare-up at Atlanta. In both cases, parents and others involved in Follow Through at the local level charged national Follow Through program officials, and particularly those involved with the evaluation, with distorting the real importance of the program. At both meetings ambiguity over the primary importance of Follow Through erupted into acrimony, and in both cases the conferences are widely recalled by participants as critical episodes in the Follow Through story.

But differences between Atlanta and Denver reveal sharp changes in the program and its evaluation between 1968 and 1972. Perhaps most obvious was the shift of alliances among parents, sponsors, and Follow Through general consultants. At Atlanta in 1968, the general consultants and parents were strongly allied, and together they charged sponsors and evaluators with insensitivity to parental and community involvement in the program. In Denver four years later, Follow Through parents were closely allied with sponsors and expressed doubts about the evaluation and the role of the general consultants.[8]

---

[7] In a memorandum dated October 30, 1972, Duane Mattheis, Deputy Commissioner of School Systems wrote to Robert Wheeler, Associate Commissioner for Elementary and Secondary Education,

> The Follow Through program was conceived and funded as a research and evaluation program to test various models and practices in compensatory education. It has always been an experimental program of fixed size and limited duration.... Follow Through is not and never was conceived of as a service program or as a separate program of compensatory education.... The scheduled phasing out of...Follow Through...means that an experimental program is being brought to an orderly conclusion. In order to avoid any misunderstanding: to state or imply that this phasing out in any way diminishes the Federal role in compensatory education or lessens support of...compensatory education...is to misrepresent the policies of USOE, the Commissioner, and the Deputy Commissioner of School Systems. (Quoted in Elmore, *Follow Through: Decisionmaking*, pp. 316-317.

The postscript to the flare-up at Denver showed how the Follow Through evaluation had changed. In late 1972, a delegation of Follow Through parents met with OE officials in Washington. Helen Testamark, leader of the parent representatives at Denver, sat on the 1972 Follow Through review panel which recommended that parents be involved in "all aspects of the Follow Through evaluation." Yet these activities bore little fruit. After 1972, there were no more national Follow Through conferences and no apparent involvement of Follow Through parent representatives in the national evaluation. So while the Atlanta episode had direct impact on the early conduct of the evaluation, the clash at Denver had no impact at all on the course of the national evaluation. The contrast may have been partly a reflection how times changed between 1968 and 1972, but also it showed how much the evaluation had been insulated from the continuing controversy over the main meaning of Follow Through.

---

[8] The parent resolution at Denver noted:

> Parents are concerned about the share of Follow Through funds that are being utilized to employ general consultants. Parents expect the general consultants to share fully their reports with PAC chairmen and to do the job which parents expect of them, other job descriptions notwithstanding. (Voices, December, 1972, p. 13.)

The 1972 reorganization had effectively tied the national evaluation to the planned variation concept of Follow Through. The main goal of the evaluation had become the assessment of program impact--the impacts of the overall program and of the different sponsor models. In focusing the evaluation in this way and narrowing the national evaluation to a standard set of outcomes, McDaniels recognized that the national evaluation would not encompass all the goals of the program nor even those of the Follow Through curriculum models. So, in setting priorities for the funding of sponsor evaluations, he gave assessment of sponsor goals not tested in the national evaluation number one priority.[9] Sponsors seemed to like the setup. One sponsor later recounted that McDaniels was:

> ...instrumental in obtaining...grant monies for each sponsor to beef up [its] evaluation...Basically Gary's point was that if the sponsors don't evaluate their work no one will, and nothing will counter the national evaluation.[1]

Plans for phasing down the Follow Through program had originally been set out in Richard Fairley's "Five Year Plan" in 1971, and it was plans for termination that ignited the controversy at Denver the following August. The phase-out plan called for limiting the program to five full cohorts of children, with the last entering kindergarten in the fall of 1972. As the start of the 1973-74 school year approached, the issue of phase-out naturally raised tensions. But by this time, the national evaluation seemed remarkably well insulated from debate. Indeed, in arguing against the phase-out one sponsor maintained that immediate phase-down of the program would have an adverse effect on the evaluation. In a memo arguing against the phase-out, the sponsor detailed a dozen ways in which the gradual phase-down might impair the integrity of the program and hence threaten the validity of the planned variation experiment:

---

[9] High/Scope Educational Research Foundation, Second Working Conference on Evaluation of Follow Through Objectives. Recommendations and Position Statements (Ypsilanti, MI: High/Scope, October, 1972), p. 17.

[1] David P. Weikart, High/Scope Educational Foundation, letter to Walt Haney Huron Institute, May 21, 1976.

> If the full collection of anticipated consequences is reasonable, no meaningful evaluation is possible after the Spring of 1974. The outcomes of Cohorts III, IV, and V will be uninterpretable. The resulting report will be so filled with caveats and qualifications as to allow only one conclusion: 'The outcomes of the Follow Through experiment in planned variation are unclear.'[2]

McDaniels had not only revamped the evaluation to correspond with the planned variation conception of Follow Through. In addition, he effectively sold it to the sponsors, or at least to some of them, and this further helped to insulate the national evaluation from the continuing controversies over the meaning of Follow Through.

In this relative isolation, Abt Associates, having won the spring 1972 contract competition, proceeded with work on the analysis of evaluation data from 1969-1972; that is, longitudinal data on Follow Through cohorts I and II, and first year data for cohort III. This phase of Abt's evaluation analysis was headed by Dr. Marvin Cline, a psychologist with previous experience in evaluating Title I compensatory education programs.

---

[2] Don Bushell, Jr., Memorandum to Garry McDaniels regarding Follow Through Termination, April 3, 1973.

A preliminary report from Abt was issued in January 1973, and drafts of the first full report appeared in late 1973. Like the early SRI evaluators the Abt approach was cautious, pointing out the numerous limitations in Follow Through evaluation data and emphasizing the very preliminary nature of analysis. "The developers of Follow Through have generally considered that the smallest time span in which sponsor impact on children, schools, teachers, and communities, can be expected to become observable is three or four years."[3]

The Abt investigators raised numerous technical issues concerning the analysis and also pursued a number of issues, like the relationship between time-of-testing and test outcomes not directly aimed at the evaluation of the Follow Through planned variations. Subsequent chapters will deal directly with technical analytical problems and with the findings of the evaluation analyses. In terms of the Follow Through evaluation story, the crucial factor of this period was that just a few months before the first full analyses based on the 1972 reorganization of the evaluation started to come off the press at Abt, there was another major shift in the leadership of the Follow Through evaluation. In the spring of 1973, McDaniels left his position as chief of the Follow Through evaluation to take a position with the newly formed National Institute of Education.

---

[3] Marvin Cline, et al., Draft Education as Experimentation: A Planned Variation Model. Volume IA: Early Effects of Follow Through. Cambridge, MA: Abt Associates, December 12, 1973, p. II-3. Emphasis in original.

Days of Reckoning and the Second Phase at Abt

> Sponsors, in particular, were attempting to maintain an atmosphere within which they could safely cooperate. Moreover, they wanted to maintain an atmosphere in which they could explore program alternatives and be genuinely innovative. If judgments of sponsors were to be based primarily on cognitive measures--reading, writing, and arithmetic, then there was the likelihood that sponsors would be afraid to be innovative; they would be more likely to play the game safely and to orient their programs so that children would show up well on the tests that were being utilized.
>
> That there was considerable emotional intensity associated with this issue should not have been surprising. After all, there was the awareness that at some point in time there would come a day of reckoning when some vague set of outsiders would sit in judgment over sponsors and projects; some sponsors and projects would continue to be supported while others would quietly be abandoned. Moreover it seemed quite clear that these judges would rely primarily on what was contained in the reports coming out of the evaluation effort.....
>
> Some of the more cynical and wordly-wise sponsors had reservations...they were skeptical that Follow Through would be able to retain control of the judgments being made, based on the SRI data. In their opinion, those individuals and agencies to whom Follow Through was responsible were likely to base their judgments almost solely on the cognitive measures that were being collected. These pessimistic opinions had the effect of an unwelcome guest at a wedding or of news of a disaster during a joyous celebration.
>
> -Account of Pajaro Dunes Conference, July 1969 by Gilbert Krulee.[4]

When McDaniels took over as chief of the Follow Through evaluation, he had moved the evaluation back under the aegis of the Follow Through program office and its director, Rosemary Wilson. At the same time, however, he effectively detached the evaluation from the service program conception of Follow Through. After McDaniels' departure, the question of where the evaluation would be housed became an issue once again.

---

[4] Krulee, Organizational Analysis, pp. 226-229.

Evaluation activities within the Office of Education had recently been centralized into an Office of Planning, Budgeting and Evaluation (OPBE) headed by Dr. John Evans. As one of the largest evaluations in OE, the Follow Through evaluation gradually was shifted into OPBE. This happened in two steps. First, Eugene Tucker, an OPBE staff member, was appointed to replace McDaniels and then, in early 1974, the Follow Through evaluation was formally transferred to OPBE.

The change in bureaucratic responsibility for the evaluation helped reawaken controversy over the relationship between the evaluation and the Follow Through program. The controversy broke open again at an annual meeting of Follow Through sponsors in October, 1974. At the conference, an OPBE representative acknowledged that the "service and R&D goals [of Follow Through] got mushed together" during "the heady period of the '60's."[5] Since Congress wanted to continue the program, he reasoned that OE ought to make the best of the situation. The situation appeared to be that Follow Through would be funded at approximately the same level as in the past--roughly $50 million a year. Since this would be too little money to expand Follow Through into a full-scale service program, OPBE advocated getting legislative recognition of Follow Through as an R&D program so that the ambiguity of Follow Through as an "experimental-cum-service program" could be ended. Specifically, the OPBE representative proposed turning Follow Through into a "constantly changing and refurbishing R&D program."[6] The Follow Through sponsors did not share this view. One contended that Follow Through had been so successful precisely because "R&D and field service aspects" had been linked together.

---

[5] The following account is based largely on a memorandum entitled "National Follow Through Sponsor Conference, Clinton, Michigan, October 29-31, 1974, Summary of Part of Discussions," Memorandum appended to High/Scope Educational Research Foundation, *Report of the Fourth Annual Follow Through Working Conference*, (Ypsilanti, MI: High/Scope, October, 1974). The High/Scope account does not identify the OPBE representative directly, but it seems clear that the individual was Evans, who was attending the conference.

The clash of views carried over to ideas about evaluation. The OPBE official contended that the evaluation was like a "horse race: a comparative evaluation of the relative effectiveness of various approaches." He acknowledged that there may be "multiple winners, in different areas," but that it is "mushy policy to say that 'all models made a fine effort so take your choice.'" According to the sponsor summary of the meeting, the OPBE official continued,

> Decisions, even though often largely political, may be influenced by evidence. Congress is often irresponsible in not facing issues. It responds to parents on the steps of the Capitol. But there is still some power in evidence. However, if we find that parents and kids are enthusiastic but the evidence is anecdotal and not persuasive, it would be irresponsible for us to tell other communities to consider the model. Arbitrary judgments will have to be made as to what evidence is compelling enough. The problem is allocation of scarce resources. We should not so enlarge the problem as to make it hopeless. Decisions are imminent...[7]

Sponsors disagreed. One protested, "I am not desirous to be a part of your horse race. It is reprehensible from an educational point of view." Another declared:

> The guy is talking about my community--but he has never been there. He says it is his program, not ours. He ignores all the pain, the struggle. He is talking about what he can do for us--and we are not asking him to do

---

[6] High/Scope, "Sponsor Conference" 1974, p. 2.

[7] High/Scope, "Sponsor Conference" 1974, p. 3.

> anything. He might be sued for libel if he says the program is not effective when we think it is.[8]

Sponsors also protested the narrow range of outcome measures used in the evaluation and stated their belief in the validity of other kinds of information, including political testimonials:

> ...It is a legitimate form of accountability for Congress to respond to people and their concerns. That is what Congress is about. "Armies of mothers might turn up--that would be an enormous contribution."[9]

The spontaneity of any army of mothers on the steps of Congress was, however, somewhat suspect in the eyes of evaluators. At the third annual sponsor conference, one working group had been devoted to the future of Follow Through and to a "discussion of methods used by sponsors to insure permanency of Follow Through and to alternatives to phase-out."[1] By 1975 the Follow Through sponsors seemed to be promoting political support for the program even more explicitly. A "directors group" meeting at the fifth annual sponsor conference came up with a list of explicit "suggestions for effective political action." Some sponsors seemed to feel that they had become unwilling participants in an evaluation horse race. As they felt this to be unfair to much of what they were trying to accomplish, they seemed to turn more directly to other avenues of influence.

During 1972-1973, the Follow Through analysis of Abt proceeded in relative insulation from the controversy over the future of Follow Through.

---

[8] High/Scope, "Sponsor Conference," 1974, p. 4.

[9] High/Scope, "Sponsor Conference," 1974, p. 4.

[1] High/Scope Educational Research Foundation, Third Annual Conference on Evaluation of Follow Through Objectives, (Ypsilanti, MI: High/Scope, October, 1973).

Some observers became worried, however, that the Abt analysis was becoming too diffuse and too little directed toward the comparative evaluation of sponsor impacts on students. Eugene Tucker, chief of the Follow Through evaluation, was concerned that the Abt analyses were becoming unwarrantedly expensive--the Abt proposal for work in 1974-75 was budgeted at roughly $1.4 million. As a result, the Follow Through evaluation analysis again went out for bids during the summer of 1974.[1] After review of alternative proposals, the evaluation contract was awarded to Abt for a second time in late 1974. The winning proposal differed substantially, however, from Abt's proposal of the previous spring. The most obvious change was in the budget. It had dropped by roughly 50 percent. The new effort at Abt was to be headed not by Cline, but instead by Richard Anderson, who had been deputy project director for the previous Abt Follow Through contracts. Also, the contract was awarded only after Abt assured the Office of Education that the second phase at Abt would "be highly centralized under a three-person management team who are concerned with addressing a prioritized list of analytic questions."[2]

---

[1] OE, Request for Proposal Analysis of Follow Through Evaluation Data Office of Planning, Budgeting and Evaluation, USOE, September, 1974.

[2] N. Ames, R. B. Anderson, M. Cline, et al., "A Proposal to Analyze Selective Sets of Data Collected on the Follow Through Program. RFP HEW/OE 75-2. Responses to Technical Questions" Cambridge, MA: Abt Associates, December 23, 1974.

As called for in the 1974 RFP, the Abt proposal focused on sponsors' effects--on children, parents and teachers. Gone was the assessment of the overall effectiveness of Follow Through. Previous evaluation reports had recounted the various features common to all Follow Through models in order to justify overall Follow Through versus Non-Follow Through comparisons,[3] but the overall Follow Through effects question was dropped in 1974 and this became a source of considerable confusion to outsiders.

In March 1973, OE officials had presented preliminary Follow Through evaluation results which showed that Follow Through pupils overall were ahead of national norms at the end of kindergarten. Moreover, the OE summary of preliminary findings reported "Follow Through's effect improves with each succeeding entering class."[4] So in 1974, Senator Norris Cotton of New Hampshire asked OE officials why they persisted in wanting to phase out the program when "some of your own evaluations indicate that Follow Through is a success." Robert Wheeler, the Associate Commissioner of Education for the Bureau of Elementary and Secondary Education responded,

> Well, Follow Through is a success as an experimental program, Mr. Chairman. In other words we think that Follow Through is doing what we want it to do now, and that is to give us some chance to look systematically at different kinds of instructional approaches, and the program does seem to be doing that for us. It is a success, but we do not feel that the experiment needs to be continued.[5]

---

[3] John Emrick et al., Interim Evaluation of the National Follow Through Program (Menlo Park, CA: SRI, February 1973) pp. 14-15; and Marvin Cline, et al., Final Report. Education as Experimentation: Evaluation of the Follow Through Planned Variation Model. Vol. IA (Cambridge, MA: Abt, March 1974), p. II-7.

[4] U.S. Congress. House of Representatives, Hearings Before a Subcommittee of the Committee on Appropriations. March 12, 1973, p. 263.

[5] U.S. Congress, Senate, Hearings Before a Subcommittee of the Committee on Appropriations, May 29, 1974, p. 2359.

Elaborating, OE's Acting Deputy Commissioner for Planning, Budgeting and Evaluation, John Evans, questioned "whether it makes sense at all to say that Follow Through is a success or failure in one single phrase that way."

> As we tried to indicate, Follow Through is made up of a different set of alternative ways of approaching compensatory education, different models, different programs. And the task and central purpose of that program, which is, as Mr. Wheeler said, primarily an experiment, is to find out which of those methods or approaches are more or less effective. The evaluation evidence we have compiled indicates just what we would expect from that kind of experiment: namely, that some of those models and approaches are very reassuringly effective, and the kinds of things we would want to see disseminated and used more broadly in compensatory education. But that same evaluation evidence indicates that a number of the other models are not successful and not effective and not the kinds of things that we would want to carry on or continue to fund or support indefinitely. That is a key point and one that I would hope that the committee would take into account in viewing this program.[6]

Senator Cotton remained perplexed:

> I have your own evaluation here put out by the Department, and I note some statements. The second paragraph starts by saying that at this stage--it is talking about Follow Through--"some very encouraging effects are being noted." And then it goes on to say "Sponsors are having some very positive effects on Follow Through children, contrasted to their comparison groups. Although the Follow Through children generally start out lower than their local comparisons on a standard test of achievement, they finished the kindergarten years slightly ahead of the comparison group."
>
> And then it goes on--I am reading special excerpts here--"However, we can conclude that when contrasting all Follow Through children with their non-Follow Through comparisons across all sponsors in all parts of the country, there emerges large achievement, motivation, and intense effects."
>
> Now, that does not sound like what you are telling me, that they have been successful in some places and a failure elsewhere. And this is put out by your Department.[7]

---

[6] U.S. Cong., Senate, Hearings, May 29, 1974, p. 2360.

[7] U.S. Cong., Senate, Hearings, May 29, 1974, p. 2361. The document to which Senator Cotton referred apparently was an OE summary of the January 1973 preliminary report by Abt.

Although the OE officials' testimony did not completely satisfy Senator Cotton, it did make clear that the overall Follow Through effects question was no longer of much interest to them. As implied in their testimony, the goal of the national evaluation in 1974, and hence the mandate for the second phase of work at Abt, was to compare the effectiveness of different Follow Through sponsor approaches.

The shift in mandate in the new Abt contract was accompanied by one other important change. Eugene Tucker, the Chief of the Follow Through Evaluation Branch, appointed Mary Kennedy, a relatively new member of his staff, to be the project officer for the Abt contract. Kennedy, who had completed her doctorate in educational psychology at Michigan State University in 1973 was to remain as Project Monitor for the Abt contract from 1974 through the end of the national evaluation effort in 1977.

The second phase at Abt began with an evaluation planning conference in Cambridge in March. At the start of the conference Kennedy set out three priorities for the evaluation effort:

- Concentrate on cohorts and streams which have completed third grade.
- More emphasis on the definition of what it means for a sponsor to be working.
- Emphasis on outcomes: what works in what situations.[8]

Kennedy's priorities met with little disagreement, but other issues provoked extended discussion.

- How can the degree of implementation of sponsors' models in different sites be assessed?
- How should third grade outcome data be analyzed?
- Should there be any follow-up testing in fourth grade after children have left the program?
- How should outcomes for sponsors be compared--against each other or against non-Follow Through outcomes?

---

[8] Abt, "Analytic Evaluation of the Follow Through Experiment. Interim Report I," 21 March 1975, Attachment B.

- How should results be presented?
- Can test data be recombined in order to derive different outcome measures appropriate to different sponsors?
- How can analyses adjust for the fact that a higher proportion of Follow Through children than comparison children attended preschool?[9]

On some of the issues general agreement was reached, but others were left open. It was agreed that: the Abt analysts would attempt to derive measures of implementation from ratings by knowledgeable observers; that there likely would be some kind of follow-up testing after children left the program; that the outcome measures ought to be used intact rather than reconstituted into new ones. Also participants agreed that more than one analytical approach ought to be used; that a multiple analyses strategy should be used so that the weaknesses of one approach could be compensated by the strengths of others. Left open were how to handle the preschool issue and how results would be presented.

In sharp contrast to many of the SRI meetings on the evaluation, the Abt conference was remarkably harmonious. The harmony stemmed from several circumstances. The evaluation design and test instruments were by then already faits accomplis. Indeed, most of the evaluation data had already been collected. A likely more important cause for the amicability was that many of the people with a direct stake in the evaluation, who in the

---

[9] Abt, "Analytic Evaluation...," 21 March 1975, Attachment B.

SRI phase of the evaluation had been party to the collaborative decisionmaking, including sponsors, general consultants, and parent representatives, did not attend the Abt conference.

The relative isolation of the Abt conference was reflective of a growing schism between the Follow Through evaluation and the Follow Through program offices in Washington. In early 1974 responsibility for the Follow Through evaluation had been officially transferred from the Follow Through Branch of OE's Bureau of School Systems to the Office of Planning, Budgeting and Evaluation. Before the end of the evaluation the bureaucratic breach between Follow Through program and evaluation offices became manifest, quite literally, in concrete. Initially the Follow Through program and evaluation offices were in adjacent connecting offices in the federal office building on the corner of 7th and D streets in Washington. However, when the offices were remodeled, the connecting doorway was taken out and the wall plastered over. And in 1976 the symbolic schism climaxed when the evaluation office moved from the 7th and D location to a building on Maryland Avenue which housed the OPBE headquarters.

The separation of the Follow Through program and evaluation efforts appeared to have no untoward effect for the Abt evaluators. In fact, the start of the second phase at Abt in 1975 was one of the smoothest periods in the history of the Follow Through evaluation. A second evaluation planning conference was held in June 1975 in Cambridge. By that time it had been decided to scrap the attempt to gather implementation ratings.[1] So the June conference focused almost entirely on how children's test data would be analyzed. It was agreed that Abt would "not make direct comparisons of sponsor Follow Through effects but rather compare FT/NFT contrasts across sponsors."[2] Also, participants agreed that three main analyses would be used: an analysis of covariance, a matching analysis and a residuals analsysis. Considerable attention was also given to how problems of preschool confounding and attrition (if and how to adjust for situations when children who took a pretest on entering Follow Through did not take the post-test), but how best to handle these issues was not completely resolved.[3]

Over the summer the Abt analysts proceeded to implement the plans agreed upon and an initial draft of the Abt III report was circulated to OE, Abt consultants and Huron staff in early fall.[4] The initial draft prompted concern over the preschool problem, the comparability of the different analyses and interaction studies which attempted to assess whether models' effects might vary for different types of children and different conditions.

---

[1] A fuller discussion of the implementation issue and why the Abt effort to assess implementation was dropped is in chapter III.

[2] Abt, "Analytic Evaluation of the National Follow Through Experiment, Interim Report V" 7 July 1975, p. 82.

[3] Again, more detail on the various analyses is given in chapter 3.

[4] Four major evaluation reports resulted from the Abt analyses of Follow Through evaluation data. The Abt I and II reports were produced in the first phase at Abt, and the Abt III and IV reports resulted from the second phase.

Over the fall the Abt team revised their volume III report and the first complete draft was finished in January 1976. The draft was issued in two parts, IIIB and IIIC. The plan was that an integrated summary of the report would be produced as volume IIIA. However, before this plan reached fruition the relative harmony of the second phase of the evaluation effort at Abt went somewhat sour.

The meat of the Abt III draft report was volume III B, which had three sections: the first consisted of introduction, plus descriptions of the evaluation sample, "measures of effectiveness" and analytical methods used in the evaluation; the second had fourteen chapters of results, one for each sponsor included in the Abt III study; and the final section was chapter 19, "Cross Sponsor Comparisons." In a spirit of openness, the Abt evaluators and OBPE staff decided that it would be appropriate to have sponsors review their respective chapters of the draft report. So the Abt evaluators sent each Follow Through sponsor a copy of the first section and the appropriate chapter of the second section.

Even though Abt did not distribute chapter 19 of the Abt III draft--the race results so to speak--the sponsor chapters roused considerable consternation. Sponsors responded with lengthy letters to Kennedy and to Tucker and phone calls and letters directly to Abt. Most sponsors objected to Abt's descriptions of their models and to the narrowness of the evaluation. Some feared that the evaluation presented an unwarrantedly negative perspective on Follow Through and several responded with technical critiques of the analyses Abt had performed.

Easily the most detailed critique came from two analysts on the staff of the High/Scope Educational Research Foundation. The High/Scope critics cited four major problems with the draft of the Abt III report:

1. It fails to explicate the methodology of data analysis--i.e., the reader doesn't know what the designs of the "adjusted gains analyses" were for specific comparisons.
2. The measures used to evaluate Follow Through programs are inadequate, but there is not discussion of their inadequacy. Rather, there is a presumption that they measure what ought to be measured for each sponsor.
3. The most important hpotheses are not tested (or are inadequately tested).
4. Serious threats to internal validity do not seem to have been overcome--i.e., analysis of covariance did not equate Follow Through and NFT samples.[5]

The High/Scope reviewers suggested that the first two problems might be overcome by rewriting and extending the Abt discussion. The last two, they maintained,"may be insuperable." In conclusion they wrote,

> *Perhaps the time has come to seriously entertain the possibility that some of the questions which Planned Variation Follow Through was intended to answer cannot be answered. If this is true (and we are inclined to believe that it is), then we should back off, re-examine the research design, and set more realistic goals for the evaluation effort.*[6]

---

[5] Judy T. McNeil and James T. Bond, "Technical Comments in Response to Abt Follow Through Reports" (Ypsilanti, MI: High/Scope Educational Research Foundation, February 24, 1976) p. 14.

[6] McNeil and Bond, "Technical Comments...," p. 16. Emphasis in original.

The sponsors were sufficiently upset over the Abt report that they organized a meeting in Washington, D.C. in April 1976, to air their grievances. The meeting was attended by representatives from Abt, OPBE and the Huron Institute, and almost all of the Follow Through sponsors. During the meeting, sponsors presented a range of criticisms; from technical to interpretative to editorial. At root, however, their concerns came down to a single question. Would it be warranted for the Office of Education to release a major evaluation report on Follow Through when the findings of the report were of doubtful validity?

Though most of the issues raised by sponsors were not resolved, the meeting closed on a compromise. Evans and Tucker agreed to postpone release of the Abt III report until further analyses were completed. Also, it was agreed that Abt would delete the concluding chapter 19 from the Abt III report in order to prevent premature conclusions being drawn from their report. Naturally, the planned volume A for the final Abt III report was also cancelled: The draft volume IIIB became final volume IIIA and draft IIIC became final IIIB.

The episode was a prime example of a Penelope ploy. Like the mythical wife of Odysseus, who wove during the day but unravelled at night, in order to postpone a decision facing her at the end of weaving, the Follow Through evaluators evidenced a recurring penchant for new analyses whenever faced with the uncomfortable call for conclusions from past analyses.[7]

---

[7] The Penelope analogy is not original. It was suggested by Ray Rhine, in "Follow Through: A Model for the Utilization of Behavioral and Social Sciences in Solving Social Problems," paper presented to the American Psychological Association Meeting, Montreal, Canada, August 27, 1973. I ought also to make clear that the penchant for postponing conclusions pending further analyses was shared by all of the Huron staff at the April 1976 meeting, including myself.

The Abt III report focused on the second evaluation cohort. The new work for Abt in 1976 was to analyze the third-grade test data for cohort III--which according to the 1972 redesign of the evaluation was to be the primary evaluation cohort. A preliminary report on cohort III findings was produced by Abt at the end of April 1976. Another interim report--a 370-page addendum to the April report, consisting mainly of data tables--was finished in July 1976. Thereafter the Abt analysts focused their energies on what was clearly to be the magnum opus of the Abt Follow Through evaluation reports. An initial draft of the Abt IV report was finished in November--and it was truly remarkable, if for nothing else its size. The five volumes of the Abt IV report, which was to be the keystone in the OPBE evaluation of Follow Through, totalled a staggering 2000 pages. This culmination of the Abt evaluation was originally scheduled for completion in December 1976, but when a second draft of the Abt IV report was finished just before Christmas 1976, the Penelope ploy replayed. Abt, and Evans of OPBE, decided that one more set of analyses--what came to be known as the bend-over-backwards analyses--ought to be carried out to insure that alternative explanation of evaluation results had been fully explored. The bend-over-backwards analyses entailed redoing the main Abt analyses after omitting various sets of projects that might be excluded for a variety of plausible reasons. So the final days of reckoning were postponed several months. The final version of the Abt IV report was approved and officially accepted by OPBE in April 1977, two months more than a decade after Lyndon Johnson proposed that the United States Congress authorize $135 million "to launch a follow-through program during the first school grades for children in areas of acute poverty."[8]

---

[8] New York Times, February 9, 1967.

II. FOLLOW THROUGH AND THE SCOPE OF THE NATIONAL EVALUATION

Even a rough history of the Follow Through program and its evaluation reveals that neither program nor evaluation has been a fixed, unchanging enterprise. Nor has there always been a complete, or even close fit between the two. Indeed, over the past ten years the Follow Through evaluation has undergone a sort of funnel vision--starting out very broad but narrowing over the years to focus on only a portion of the program and on a limited set of the questions that have been asked about it.

The purpose of this chapter is to examine this focus, to show what's been left out and to identify questions the evaluation has addressed in light of the questions that might have been addressed. Also, we try to begin explaining why the evaluation has narrowed to focus on one particular set of questions.

The chapter has four parts. The first describes the official definition of Follow Through and outlines the scope of the program in terms of the numbers of sponsors and projects in the program and the waves or cohorts of children who have participated in it. Against this backdrop, subsequent parts compare the national evaluation to the range of evaluation criteria suggested for the program, to the numbers of sponsors, sites and cohorts in the program, and finally to different types of evaluation questions that might be posed about the program.

Attention is limited here to the national Follow Through evaluation and mainly to what we term the major national Follow Through evaluation reports. These nine evaluation studies are listed in Table 2, together with the brief names by which they are called hereafter.

Table 2: Major National Follow Through Evaluation Studies

Brief Name	Full Reference
Emrick, Interim Evaluation	John A. Emrick, Philip Sorenson, and Marian S. Stearns. Interim Evaluation of the National Follow Through Program 1969-1971. A Technical Report. Menlo Park, CA: Stanford Research Institute, February, 1973. 325 pp. ED 086 371.
Soar, Classroom Process	Robert S. Soar. Final Report. Follow Through Classroom Process Measurement and Pupil Growth (1970-71). Gainesville, FL: College of Education, University of Florida, June 1973, 231 pp. Plus Appendix E. Sponsor Differences in Classroom Process, June 1976, 29 pp.
Stallings, Classroom Observation 1971-72	Jane A. Stallings, Follow Through Program Classroom Observation Evaluation 1971-72. Menlo Park, CA: Stanford Research Institute, August 1973. 547 pp. ED 085 100.
Cline, Abt I	Marvin G. Cline, et al. Final Report. Education as Experimentation: Evaluation of the Follow Through Planned Variation Model. Volume IA: Early Effects of Follow Through and Volume IB: Monographs. Cambridge, MA: Abt Associates, March 1, 1974. Altogether ca. 800 pp. ED 094 890 & ED 094 891.
Stallings, Classroom Observation 1972-73	Jane A. Stallings, David H. Kaskowitz, and others. Follow Through Classroom Observation Evaluation 1972-73. Menlo Park, CA: Stanford Research Institute, August 1974. ca. 600 pp. ED 104 891.
Cline, Abt II	Marvin G. Cline, et al. Final Report. Education as Experimentation: Evaluation of the Follow Through Planned Variation Model. Volume IIA: Two-Year Effects of Follow Through and Volume IIB: Appendices and Monographs. Cambridge, MA: Abt Associates, May 1, 1975. Altogether ca. 800 pp. ED 108 768 & ED 108 769.
Stebbins, Abt III	Linda Stebbins, et al. Education as Experimentation: A Planned Variation Model. Volume IIIA: Findings Cohort II; Interim Findings Cohort III and Volume IIIB: Appendices. Cambridge, MA: Abt Associates, July 1975. Altogether ca 835 pp.
Abt IV **	Linda Stebbins, et al. Education as Experimentation: A Planned Variation Model. Cambridge, MA: Abt Associates, April 1977.
Abt IVA	L. Stebbins, R. St. Pierre, E. Proper, R. Anderson and T. Cerva. Volume IVA. An Evaluation of Follow Through. 277 pp.
Abt IVB	G. Bock, L. Stebbins, et al. Volume IVB. Effects of Follow Through Models. 198 pp.
Abt IVC	Volume IVC. Appendices. Part 1 and Part 2. 911 pp.
Abt IVD	T. Ferb, J. Larson, D. Napior, et al., Part 1: A Longitudinal Study of Follow Through. Bernadette Nelson, et al., Part 2: An Academic Risk Study. 208 pp.
Abt NFT	J. Molitor, M. Watkins, D. Napior, et al. Non-Follow Through Study. 587 pp.
RMC Cost Study	Donald Rogers, et al. Final Report. A Cost Analysis of Follow Through Projects. April 1977. ca. 100 pp. plus 8 appendices, ca. 200 pp.

*Summaries of these major national evaluation reports are given in the appendix to this volume.

**Abt's volumes IVA and IVB became volumes IIA and IIB in OE's final series entitled The Follow Through Planned Variation Experiment. Also, two additional volumes in the Abt IV series are scheduled for release in September 1977. These volumes, IVE and IVF in the Abt series, will become IIC and IIC appendix in the OE final report.

We focus on these nine studies for several reasons. They are all relatively wide in scope; each includes information on at least a dozen projects associated with at least eight different sponsors. All of the major studies were carried out under contracts with the U.S. Office of Education and have been officially accepted by the USOE Follow Through evaluation branch. Together these reports represent the bulk of the national Follow Through evaluation effort.[1]

Several dozen other reports have been prepared for the Follow Through evaluation branch, but are not systematically included here. Many of these are draft or progress reports, or treat special topics tangential to the main evaluation effort. Where appropriate these reports are mentioned in this review, and all that we have been able to locate are included in the bibliography, but the focus of the review is the set of nine major studies.

What is Follow Through?

Official statements of aspirations for Follow Through exist in numerous forms, but no single definition of the program seems ever to have been universally accepted. The original 1967 legislation called Follow Through a program for "children in kindergarten or elementary school who were previously enrolled in Head Start or similar programs and designed to provide comprehensive services and parent participation activities [to] aid in the continued development of children to their full potential."[2]

---

[1] The Pittsburgh national evaluation study for 1967-68 is not included as a major report because it predates the planned variation phase of Follow Through. Also, it should be noted that SRI investigators pursued a variety of issues in the years 1969-72 which never were described in any final reports. Several of these efforts will be discussed in chapter III.

[2] Economic Opportunity Act of 1964, Public Law No. 88-452, 78 Stat. 508, as amended in Public Law No. 90-222, tot. I, sec. 104, 81 Stat. 698, 42 U.S.C. 2809 (a) (2) 1967.

The program was reauthorized by Congress in the Economic Opportunity Amendments of 1972.[3] Two years later, the Community Services Act of 1974 re-extended authorization for the program and elaborated on the developmental and evaluation aspects of the program:

> Sec. 553. (a) In conjunction with other activities authorized by this part, the Secretary may--
>
> (1) provide financial assistance by contract or otherwise, for research, demonstration, or pilot projects conducted by public or private agencies which are designed to test or assist in the development of new approaches or methods that will aid in overcoming special problems or otherwise in furthering the purposes of this part; [and]
>
> (2) provide, directly or through grants or contracts, for the continuing evaluation of projects assisted under this part, including evaluations that describe and measure the impact of such projects, their effectiveness in achieving stated goals, their impact on related programs, and their structure and mechanisms for delivery of services, including, where appropriate, comparisons with appropriate control groups composed of persons who have not participated in such projects, which evaluations shall be conducted by persons not directly involved in the administration of the project evaluated.[4]

Guidelines for the Follow Through program were issued in February 1969 in the form of a draft program manual--"for information purposes only." Final program rules and regulations were not published until April 1975, nearly eight years after the first children entered the program. According to Robert Egbert, the delay in finalizing the regulations stemmed from division of responsibility for Follow Through between OEO and OE, and lack of HEW commitment to the program.[5] A more immediate cause for delay was controversy over the main meaning and hence the official description of the program.

---

[3] Economic Opportunity Amendments of 1972, Public Law No. 92-424, secs. 8-11, 86 Stat. 690, 691, 42 USC 2809 (a) (2) (1972).

[4] Community Services Act of 1974, P.L. 93-644, Title V, Part B; 42 USC 2929, known as "The Head Start-Follow Through Act."

[5] R. Egbert, personal communication to W. Haney, January 1977.

A notice of proposed rule making for Follow Through was published in the Federal Register in early March 1974. In accordance with Education Amendments of 1972, public comments on the proposed regulations were solicited and an open meeting on the regulations was held in Washington on March 28, 1974. Around 800 written comments were received, and as the Register recounts,

> Virtually all the comments received [at the public meeting or in writing] expressed opposition to the characterization of Follow Through...as "experimental" and expressed opposition to the contemplated phase-out of Follow Through...[6]

In reponse to the criticism, the language requiring the phase-out was deleted, although the reference to Follow Through as an "experimental" program was retained. So, on April 21, 1975, the Register of the National Archives of the United States officially announced:

> The Follow Through Program implemented by these regulations is an experimental community services program designed to assist, in a research setting, the overall development of children enrolled in kindergarten through third grade from low-income families, and to amplify the educational gains made by such children in Head Start and other similar quality preschool programs by (a) implementing innovative educational approaches, (b) providing comprehensive services and special activities in the areas of physical and mental health, social services, nutrition, and such other areas which supplement basic services already available within the school system, (c) conducting the program in a context of effective community service and parental involvement, and (d) providing documentation on those models which are found to be effective.[7]

---

[6] Federal Register, Vol. 40, No. 77 (April 21, 1975), part II, p. 17712.

[7] Office of Education, Dept. of Health, Education and Welfare, "Follow Through Program Rules and Regulations," Federal Register, Vol. 40, No. 77 (April 21, 1975) 17712-17721. Interim final regulations were issued in 1974 (Federal Register, Vol. 39, No. 121, June 21, 1974, 22342-2235), but the final final regulations were not issued until 1975.

The Register also described the planned variation aspect of Follow Through and the evaluation of the variations:

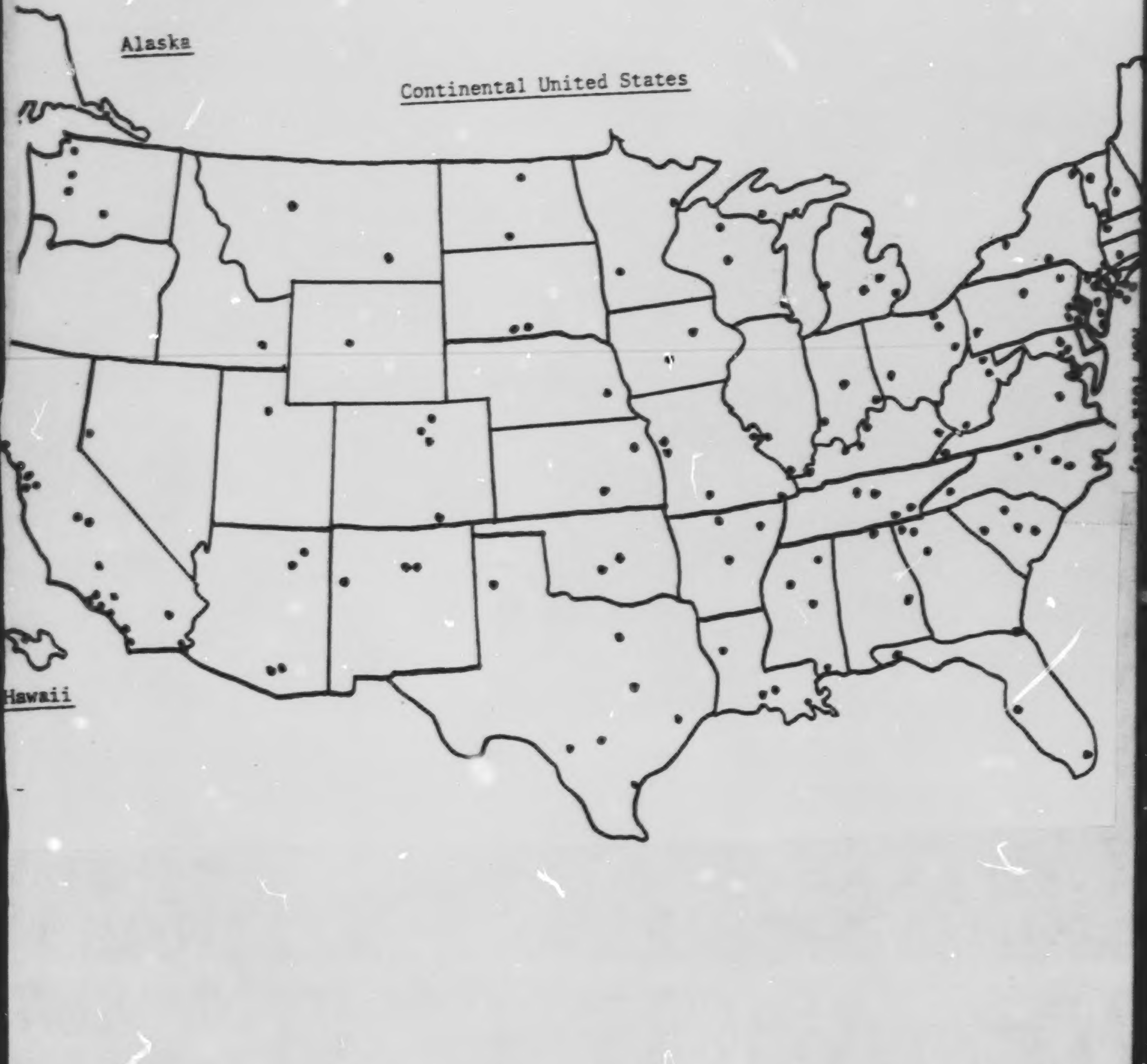
> (a) Follow Through project grants are made to local educational agencies and other public or non-profit private agencies, organizations, or institutions in order to explore the effects of a number of promising approaches to the education of children from low-income families in the early elementary grades. Most grantees must agree to carry out the project in cooperation with project sponsors who have developed such approaches in affiliation with the U.S. Office of Education.
>
> (b) In order to assess the effectiveness of each approach, evaluation data is being collected by the U.S. Office of Education. Because the collection of data will continue for a number of years, project and sponsor grantees are required to work together in the development and implementation of the sponsor's approach for the period of their participation in the Follow Through program except if the Commissioner determines otherwise.[8]

The regulations describing Follow Through became effective in June 1975. By that time there were around 170 Follow Through projects, at least one in every state in the Union. The distribution of projects is shown in Figure 1. Most projects are associated with one or more Follow Through model sponsors. Altogether there have been 23 different sponsors. One dropped out in 1970 and two more in 1975, so by 1976 there were only twenty. Also, some Follow Through projects are not associated with any sponsor because they are self-sponsored or parent-implemented. The different types of Follow Through sponsorship are summarized briefly in Figure 2. Roughly 70 percent of the Follow Through projects are associated with one of twelve sponsor approaches initially included in the planned variation phase of the project in 1968.

---

[8] OE, "Rules and Regulations," Federal Register, Vol. 40, No. 77, 17714. In addition to the change regarding phase-out, public commentary prompted changes in the language of the regulations so as to stress "the importance of the type and extent of parental participation" in Follow Through.

Figure 1: Geographical Distribution of Follow Through Projects



As a quick perusal of Figure 2 reveals, Follow Through sponsors' models vary widely; from programs based on behavior modification principles to approaches stressing different kinds of open classrooms. Some sponsors want to teach basic skills, to do better what schooling traditionally has done. Others are out to reform schooling, to make it more humane and flexible. Most sponsors focus on school curriculum but others stress parental involvement and have little to say directly about classroom processes.

Several different schemes for grouping sponsors have been suggested, variously relying upon sponsors' institutional affiliations, educational goals, and philosophies, or the degree of structure in their models. However, different categorization schemes include different sets of Follow Through sponsors; and different categorizations have been used in different national evaluation studies.[9] Therefore, in the remainder of this report we generally refer directly to individual sponsors rather than to types of sponsors.

Since the planned variation phase of Follow Through began in 1968, a number of waves or cohorts of students have participated in Follow Through. Within each cohort there are two types or streams of children--those who enter kindergarten and those who enter directly into first grade because their schools have no kindergarten. The Follow Through projects which do not have kindergarten, and hence are termed entering-first (EF) sites, are located largely in the southern part of the United States.

---

[9] For example, among major national evaluation studies, the Abt III report categorizes sponsors by degree of structure and theoretical emphasis (Abt IIIA, p. A-22) and Stallings following Maccoby and Zellner classifies them according to program orientation (Stallings, 1971-72, pp. 9-10). The Abt IV report categorizes the 17 sponsors encompassed in that report along four dimensions: emphasis on learning domain (basic skills, cognitive/conceptual, cognitive/affective); degree of structure (high, medium, low); emphasis on home learning environment (low, moderate, high); utilization of parents as teaching agents (low, moderate, high). Abt IVA, pp. 22-27.

Figure 2: Follow Through Sponsors

Title of Sponsor's Model or Other Category[1] (Sponsor Code Number)[2]	Sponsor institutional affiliation and director(s)[3]	Curriculum Aspects of Model[4]	First School Year in Program[5] (Last Year in Program)	Number of Projects or Site
Self-sponsored projects (01)		Sites which choose to implement local FT programs without affiliating with curriculum model sponsor.	1967-68	13
Responsive Education Program (02)	Far West Laboratory for Educational Research and Development [Glen Nimnicht] Denis Thoms	Develop problem-solving skills, sensory discrimination, self-confidence with self-paced instruction.	1968-69	15
Tucson Early Education Model (TEEM) (03)	University of Arizona [Marie Hughes, Ronald Henderson] Joseph Fillerup	Open Education. Concentration on broad intellectual skills, positive attitude toward learning, positive reinforcement of child's interests.	1968-69	20
Bank Street Model (05)	Bank Street College of Education [Elizabeth Gilkeson] Lorraine Smithberg	Develop positive self-image, creativity, coping skills, use of language to formulate ideas and express feelings.	1968-69	14
Mathemagenic Activities Program (06)	University of Georgia [Charles Smock] [Gilbert Shearron] Horace Hawn	Learning-by-doing, positive reinforcement. Interaction with environment to develop initiative and decision-making by the child.	1968-69	7
Direct Instruction Model (07)	[University of Illinois] University of Oregon Siegfried Engelmann Wesley Becker	Basic skills in reading, arithmetic, language. Highly structure drill; reward, praise for reinforcement.	1968-69	20
Behavior Analysis Approach (08)	University of Kansas [Donald Bushell] Gene Ramp	Behavior analysis (modification), uses token economy to reward individual achievement; programmed materials emphasizing reading, writing, and mathematical skills.	1968-69	13
Cognitively Oriented Curriculum Model (09)	[Eastern Michigan University] High/Scope Foundation David Weikart	Piagetian theory, emphasizing stages of intellectual development; detailed plans for classroom organization; home teaching that involves parents.	1968-69	10
Florida Parent Education Follow Through Model (10)	University of Florida [Ira Gordon] Ira Gordon, Gordon Greenwood, William Breivogel	Parent education, aides work with parents in home and parents teach children; curriculum content varies with individual needs.	1968-69	11
EDC Open Education Model (11)	Education Development Center [David Armington] [David Hawkins] [George Hein] Grace Hilliard	Open classrooms approach aimed at developing self-respect, imagination, persistence, openness to change, critical abilities.	1968-69	10
Individualized Early Learning Program (12)	University of Pittsburgh [Lauren Resnick, Warre Shepler] R. Eichelberger, Elizabeth Boston	Individually prescribed instruction; programmed learning materials; emphasis on academic skills and positive reinforcement.	1968-69	7
Interdependent Learning Model (13)	[New York University] [City University of New York] Fordham University [Lassar Gotkin] [Donald Wolff, Jack Victor] Harold Freeman	Small group instructional games, children teaching children, develop language, reading, logic, mathematical ability.	1968-69	3

FIGURE 2: FOLLOW THROUGH SPONSORS (CONTINUED)

Title of Sponsor's Model or Other Category (Sponsor Code Number)		Sponsor institutional affiliation and director(s)	Synopsis of Model	First School Year in Program[5] (Last Year in Program)	Number of Projects or Sites
Language Development (Bilingual Education) Approach	(14)	Southwest Educational Development Laboratory [Juan Lujan] [Don Williams] Preston Kronkosky	Bilingual language development for Spanish-speaking children. Positive emphasis on child's native language and culture.	1968-69	
Parent Implemented	(15)		Not necessarily affiliated with any sponsor; projects in which Policy Advisory Committees, including parents tend to play larger role in policymaking than those sites classified as self-sponsored	1968-69	14
Hampton Institute Non-graded Model	(17)	Hampton Institute [Martha Dawson] Mary T. Christian	Non-graded, flexible diagnostic and prescriptive teaching keyed to ethnic background of children.	1969-70	6
Culturally Linguistic Approach	(18)	Northeastern Illinois University [Nancy Arnez] [Clara Holton] Donald Linder	Use of non-standard English spoken by children as basis for development of patterns of thought and language.	1969-70	5
Parent Supported Application of the Behavior Oriented Prescriptive Teaching Approach	(19)	[State College of Arkansas] Georgia State University Walter Hodges	Parents as tutors, individually prescribed learning, to meet operationally described learning needs of children.	1969-70	2
[Responsive Environments Model] [Programmed Instruction Model] Personalized Instruction Model	(20)	[Responsive Environments Corporation] Prentice-Hall Developmental Learning Centers [Lorie Caudle] [Marty Dummet] Marilyn Smetana	Provide highly individualized program to allow children to progress at own pace by using teaching machines, programmed learning.	1969-70	1
Home-School Partnership Model: A Motivational Approach	(21)	[Southern University] Clark College [Edward E. Johnson, Ann Price] Johnnie Follins, Jean Humphrey	Parent involvement in teaching and governance of schools. Make school setting more relevant to home and children.	1969-70	2
California Process Model	(22)	California State Department of Education [Ruth Holloway] James Jordan	Coordination of home, school and community in promoting education. Joint determination of learning goals.	1969-70 (1975)	6
New School Approach	(23)	University of North Dakota Vito Perrone	Open classroom with children responsible for their own learning; teachers, aides, parents and older students assist.	1970-71	4
Parent Implementation Educational Model	(24)	AFRAM Associates Preston Wilcox	Parent involvement in selection and development of instructional program. Development of parent skills in promoting educational programs to serve needs of community.	1970-71 (1975)	9
Culturally Democratic Learning Environments	(26)	[University of California - Riverside] University of California - Santa Cruz Manuel Ramirez III	Bilingual, bicultural approach for teaching concepts in math and science and development both in reading skills and oral expression.	1971-72	1
Role Trade Model	(27)	Western Behavioral Sciences Institute [Stanley Crockett] Wayman Crow	Role-trade, allowing parents, children, and community participants to exchange roles.	1971-72	1

(Notes to Table 2 on following page.

Total affiliations = 199
Total projects without double counting = 182

Notes to Figure 2

1. This column lists 24 types of sponsorship. Twenty-two are sponsor model titles. The self-sponsored category refers to projects which have no affiliation with other sponsors, but which instead act as their own curriculum model sponsor. The parent-implemented category refers to projects in which parents play an especially large role in policymaking for the local project. Most of the parent-implemented projects are also affiliated with some other curriculum sponsor.

   Titles of sponsors' models vary somewhat in different descriptions. For example, the Personalized Instruction Model has previously been called the Programmed Instruction Model and before that the Responsive Environment Model. The California Process and the Afram Parent Implementation Models were not included in Project Follow Through after 1974-75.

2. Sponsor code numbers were assigned by SRI for the purposes of data-gathering. The numbers are not consecutive because code number 25 was not used, number 16 was temporarily used to indicate projects undecided as to sponsorship, and 04 was originally used to indicate the DARCEE model sponsored by George Peabody College, which "had by mutual consent terminated its association with Follow Through as of the close of 1969-70 school year." Egbert, "Follow Through," p. 14.

3. The institutional affiliations and the directors of the sponsor curriculum models have in a few cases changed since the start of planned variation. Insofar as is known, former directors and institutional affiliations are shown in brackets.

4. These brief descriptions provide only a very rough idea of sponsors' approaches. Fuller descriptions are available in Volume III of OE's final report on the Follow Through national evaluation.

5. Thirteen sponsors joined the Follow Through program in 1968, seven in 1969, and three more in 1970-71. Two sponsors dropped out in 1975.

6. It is difficult to summarize how many projects are affiliated with each sponsor; some are affiliated with more than one sponsor, and a few affiliations have been terminated during the course of the program. In this table, several projects affiliated with more than one sponsor have been double-counted--once in each category. Hence the total number of projects listed here is 199 while the total number of different projects which have been in Follow Through is only 182.

   The terms "site" and "project" seem to have been used largely interchangeably in the Follow Through evaluation. "Project" seems to be favored by SRI evaluators while "site" is the term more often employed by Abt analysts. Outside of the evaluation, the term "project" has been used much more widely than "site." Project generally refers to one or more schools affiliated with a Follow Through model sponsor (or self-sponsored or parent-implemented) in a particular location. For a listing of all projects see SRI, Follow Through Evaluation Data Bank File Description All Years Beginning 1972-73, Part 2, Appendix A. This appendix does not list one project (#12.02) given by Jane David's Final Report on Sample Selection (Cambridge, MA: Huron Institute, June 30, 1973), Attachment 8.

The first group of children entered the Follow Through planned variation experiment in 1968-69. As was mentioned in Chapter I, it was decided in 1970 that this group of children would be considered an implementation cohort and hence not subject to evaluation. Thus the cohort of children entering Follow Through in 1969 came to be known as cohort I. After the 1972 re-organization of the evaluation, cohort III, which entered the program in 1971, became the primary evaluation cohort. The temporal relations among the Follow Through cohorts and streams are depicted in Figure 3. Other general information on the scope of the program was given in Table 1, p. 21 .

From these data, it is easy to see that Follow Through has been a program of tremendous scope. It has spanned nearly a decade, includes some 170 projects across the United States, a wide range of services, and some twenty different sponsors' model approaches. All of this has been complicated by the ongoing controversy over the main meaning of Follow Through--whether it is primarily a service program, a social action program, a planned variation experiment, or a forum for the development of new curriculum model approaches. Within this scope and without detailing more of this controversy, the next section begins to sort program elements covered in Follow Through evaluation reports from those which have received little attention.

Evaluation and the Scope of the Program

The official rules and regulations for the Follow Through program suggest several kinds of criteria by which program effectiveness might be evaluated:

FIGURE 3: FOLLOW THROUGH COHORTS AND

FT Cohorts*	Cohort Stream	1967-68	1968-69	1969-70	1970-71	1971-72
Pilot		K	1	2	3	
		1	2	3		
Implementation			K	1	2	3
			1	2	3	
I	I EK			K	1	2
	I EF			1	2	3
II	II EK				K	1
	II EF				1	2
III	III EK					K
	III EF					1
IV	III EK					
	III EF					

*Each FT cohort contains two streams, an entering kindergarten (EK) stream and

STREAMS

1972-73	1973-74	1974-75	1975-76
3			
2 3	3		
1 2	2 3	3	
K 1	1 2	2 3	3

an entering first grade (EF) stream.

(1) Comparisons of cognitive and affective development of the children served with non-Follow Through children;

(2) Comparisons of the cognitive and affective development of Follow Through children who participated in Head Start and other quality preschool programs with the cognitive and affective development of Follow Through children who did not;

(3) Extent of parent involvement in the project, including participation in the decision-making process and participation in the classroom as paid employees or volunteers, and the effect of the project upon parental attitudes concerning the school and education in general;

(4) Extent of medical and dental care, of psychological services, and of social services available to low-income Follow Through children;

(5) Evidence of changes in school programs and in institutional rules and practices which increase the responsiveness of the educational system to low-income children and their parents;

(6) Evidence of positive changes in attitudes of program participants twoards themselves and the community; and

(7) Evidence of coordination of community services (e.g., medical services) with the participating schools and their responsiveness to the needs of low-income persons.[1]

A checklist of these criteria is given in Table 3. The checklist shows which criteria are addressed in which of the major, national Follow Through reports. The X's indicate criteria which have been directly addressed in the major evaluation reports, while question marks indicate criteria which are indirectly or partially reflected in the evaluation reports. The patterns of X's, ?'s and blanks reveal a clear focus to the overall national evaluation. Of the evaluation criteria suggested

---

[1] OE, DHEW, "Follow Through Program Rules and Regulations," Federal Register, Vol. 40, No. 77 (April 21, 1975), 17716-17.

Table 3: Focus of Follow Through Evaluation Reports*

Suggested Criteria for Follow Through Evaluation	Major National Evaluation Reports (Date of Completion)									Other National Evaluation Reports
	Emrick (2/73)	Soar (6/73)	Stallings 1971-1972 (8/73)	Cline Abt I (3/74)	Stallings 1972-1973 (8/74)	Cline Abt II (5/75)	Stebbins Abt III (7/76)	Stebbins Abt IV (4/77)	RMC	
1) Cognitive & affective development; FT pupils vs. NFT.	X	X	X	X	X	X	X	X	(Costs of Models)	
2) Cognitive & affective development; FT pupils w/HS vs. FT w/o HS.	?			X		X		X		
3) Parent involvement; participation & effects of project on parent attitudes.	?			?		?	?	?		SRI, Draft Appendix E: FT Community Studies (1972)
4) Comprehensive services; medical & dental care, etc.									(Costs)	Sullivan, Health Care Delivery in Proj. FT (1970)
5) Changes in school programs and institutional rules & practices.		?	?	?	?					Hargroves, Pupil Assignment (1973)
6) Changes in attitudes of program participants.	?	?	?	?	?	?	?	?		
7) Coordination of community services with participating schools.										Krulee, Organizational Analysis (1973)

* The criteria for the evaluation of FT were suggested in the FT rules and regulations, Federal Register, Vol 40, No. 77 (April 21, 1975) pp. 17716-17. The seven suggested criteria are given in full in the text above. The X's in the checklist indicate criteria that have been directly addressed in major evaluation reports and question marks (?) indicate ones that are used only indirectly. Note that the sixth criteria overlaps with the first three.

** Reports listed here refer to auxiliary evaluation studies pertinent to the different suggested criteria. Such auxiliary or supplementary studies were commissioned during both the Egbert-Snyder and the Wilson-McDaniels phases of the program. Many but not all of these auxiliary studies are included in the bibliography.

in the Follow Through program rules and regulations, the national evaluation has focused most directly on the comparison of cognitive and affective development of Follow Through pupils with that of non-Follow Through pupils. The evaluation has given least attention to the comprehensive services provided through the Follow Through program and to the coordination of services with those of the school and community. Other topics listed in Table 2 have been addressed in the national evaluation reports either indirectly or partially. For example, none of the major national evaluation reports directly examines whether Follow Through has promoted changes in "school programs and in institutional rules and practices," but this topic is indirectly addressed in the Stallings and Soar reports and in the first Abt report in a section on implementation. The other three kinds of criteria mentioned in the Follow Through rules and regulations are likewise addressed only indirectly.

Topics given little consideration in the major national evaluation reports have in some cases been treated in early draft reports which never reached final form, or in supplementary, national evaluation reports.[2] The supplementary studies are far smaller--in scope, in duration, and in funding--than any of the major national reports. Hence, they have tended not to address any of these topics in great depth or in many settings.

The pattern shown in Table 3 reflects the history of the Follow Through evaluation and what has caused continuing controversy over the evaluation effort: The major share of the national Follow Through evaluation has focused on a limited part of the program, and has given relatively little heed to elements of the program which some people believe vital to its success.

---

[2] For example, see the annotation in the bibliography on Sullivan's Health Care Delivery in Project Follow Through.

The locus of attention has been the program's impact on children's cognitive and affective development or more specifically, on children's achievement and attitude test scores. The focus on children instead of on parents, teachers, or other Follow Through participants mirrors the ultimate goal of the effort--promoting the development of low-income children to their full potential--but it largely ignores the social action, service delivery and R&D conceptions of the program.

The focus on children's test scores reflects mainly the planned variation conception of Follow Through. Latching the evaluation to the planned variation conception of Follow Through resulted partly from necessity. In its early phases, the evaluation was so bounced and stretched between different conceptions that much of it miscarried. The draft of SRI's 1969-71 report issued during the summer of 1972, including four annexes and six appendices, ran to some 800 pages in length.[3] By February 1973, when the report reached final form, it had shrunk to 325 pages and completely omitted substantial portions of the draft report dealing with pupil tests, parent interviews, teacher questionnaires and community studies. One SRI staffer later recounted that as attention shifted to the Abt Follow Through analyses, these portions of the draft report "simply fell through the cracks [at OE]."[4]

---

[3] Stanford Research Institute, Draft Evaluation of the National Follow Through Program 1969-71 (Menlo Park, CA: SRI, August 1972). Annexes I-IV were bound together in a separate volume, and each of the six appendices was bound separately:

A. Follow Through Program Sponsors
B. Follow Through Classroom Observations
C. Follow Through Pupil Tests, Parent Interviews, and Teacher Questionnair
D. Follow Through Project Descriptions
E. Follow Through Community Studies
F. Administrative History of the Follow Through Evaluation, 1968-72.

McDaniels' revamping of the evaluation, latching it to the planned variation conception of Follow Through, was partly the cause of the withering away of much of the SRI draft report. Also, however, the overhaul was an effect of the diffuseness of the evaluation effort at SRI. McDaniels and his associates adopted the planned variation conception of Follow Through partly as a means for forcing some shape out of the amorphousness of previous evaluation efforts.

Necessity does not, however, completely explain the focus on planned variation instead of on, say, the service conception. The narrowing view of the evaluation on the planned variations also derived from the intellectual origins of the planned variation notion. During the late 1960's, there was growing dissatisfaction with large-scale evaluations which lumped programs together from a wide range of settings, attempted to calculate an overall effect, and almost inevitably came up with negative results, or conclusions of "no significant effects." The view was aptly expressed by Urie Bronfenbrenner in 1969 in comments on the disappointing results of the Westinghouse-Ohio evaluation of Head Start:

> I sat around a table with several colleagues, two years ago now, and we said explicitly, "If this study is carried out as it is designed, the results will be negative and we won't know whether they are really negative. There may actually be positive findings there that won't show....[It] is not the kind of research design that will yield us the information we need to know. It is guaranteed to give us zero results,

---

[4] Telephone conversation, February 1976, with SRI staff member.

> because when you put everything in the same box, the good programs will cancel out the bad programs, and you will get, you know, a great nothing.[5]

The notion of planned variation was motivated by the desire to avoid this problem, to preclude lumping disparate programs together "in the same box." So after the Follow Through evaluation crises of 1971, when it came time to revamp and reform the evaluation effort, it seemed natural to latch it to the planned variation conception of Follow Through. After all, it might have been argued, if the unplanned variations in comprehensive services or in parental involvement were to be lumped together in the same evaluation, it is highly probable that the result would be "a great nothing."[6]

Evaluation and the Scope of Planned Variation

Whatever the cause, it is clear that the national evaluation has focused largely on Follow Through as a planned variation experiment. In delineating the scope of the national evaluation, the natural next question is how does the focus of the evaluation fit the scope of the planned variations?

A rough answer to this question is shown in Table 4. The chart indicates which of the 22 Follow Through sponsors are included in each of the major Follow Through evaluation reports. Ten of the Follow Through sponsors were not included in any of the first six major national studies, but the last three studies--the Abt III, the Abt IV and the RMC reports--included information on a wider range of sponsors. All but the Afram parent implementation approach have been included in at least one major report. The Afram approach received scant attention for several reasons: it was one of the late-starting sponsors; it focused not on a particular curriculum approach but on parent involvement; and the Afram sponsorship was not continued after 1974-75. Overall only six sponsors are included in all nine major studies.

---

[5] U.S. Congress, House of Representatives, Hearings Before the Select Subcommittee on Education of the Committee on Education and Labor, House of Representatives on H.R. 13520 (Washington, D.C.: GPO, 1970), p. 157.

[6] From his study of health components of Project Follow Through, Thomas Sullivan concluded "because of the variances in local circumstances from community to community, achievements in one community would appear as nothing in another." Health Care Delivery in Project Follow Through (Cambridge, MA: Bio-dynamics, November 16, 1970), p. 1.

This view provides only a rough image of the fit between the evaluation and the scope of the planned variations. To be more precise, we have to consider this issue at the level of projects or sites.[7] Thirteen sponsors are included in five or more of the major, national evaluation studies encompassing a heavy majority of Follow Through projects, but it is important to note that not all of the sites associated with any one of these thirteen has been included in any of the major national evaluations. The distribution of Follow Through projects included in each of the major, national evaluation reports is depicted in Table 5. While the pattern varies somewhat across reports and among projects associated with different sponsors, the general pattern is clear. Most reports deal with less than half the projects associated with any sponsor.

---

[7] See note 6, Figure 2 for a description of how these two terms differ.

Table 4: Sponsors Included in Major Evaluation Reports*

		MAJOR EVALUATION REPORTS								
SPONSOR	SPONSOR CODE	Emrick	Soar	Stallings 1971-72	Cline Abt I	Stallings 1972-73	Cline Abt II	Stebbins Abt III	Stebbins Abt IV	RMC
Self-sponsored	01	X						X	X	
Responsive Education	02	X	X	X	X	X	X	X	X	X
TEEM	03	X	X	X	X	X	X	X	X	X
Bank Street	05	X	X	X	X	X	X	X	X	X
Mathemagenic	06	X		X	X**				X	X
Oregon	07	X	X	X	X	X	X	X	X	X
Behavior Analysis	08	X	X	X	X	X	X	X	X	X
Cognitive Curriculum	09	X		X	X	X	X	X	X	X
Florida Parent Education	10	X	X	X	X		X	X	X	X
EDC Open Education	11	X	X	X	X	X	X	X	X	X
Individ. Early Learning	12			X	X		X	X****	X	X
Interdependent Learning	13	X	X	X				X	X	X
Language Development	14	X		X	X		X	X	X	X
Parent Implemented	15	X								
Hampton Non-graded	17									X
Culturally Linguistic Approach	18								X	X
Parent Supported	19									X
Responsive Environments	20									X
Home-School Partnership	21							X	X	X
California Process Model***	22							X	X	
New School Approach	23								X	X
Afram Parent Implementation***	24									
Cultural Democratic Learning	26									X
Role Trade Model	27									X
TOTALS		13	8	12	11	7	10	14	17	20

*A sponsor is marked as included in a report if one or more sites associated with it is included in the report. Each of the major national evaluation reports includes several analyses. If one of a sponsor's sites is included in only one of a report's analyses, it is marked as being included in that report.

**Sponsor 6 was included in the Abt I Cohort I and Multiple Cohort studies but not in the Abt I main analyses.

***The California Process Model and the Afram Parent Implementation Model were not included in Project Follow Through after 1974-75.

****The Individualized Early Learning Program was included in the Abt III interim results for cohort III but not in the cohort II results which were the focus of the Abt III report.

Table 5: Projects Included in Major Evaluation Reports*

SPONSOR	SPON. CODE	Total**	Total as of July'76	PROJECTS ASSOCIATED WITH EACH SPONSOR INCLUDED IN DIFFERENT REPORTS								
				Emrick	Soar	Stallings 1971-72	Cline, Abt I	Stallings 1972-73	Cline Abt II	Stebbins Abt III	Stebbins Abt IV	RMC
Self-sponsored	01	13	14	5						5	5	
Respon. Educ.	02	15	14	3	5	1	5	5	6	7	11	3
TEEM	03	20	19	4	5	1	6	6	6	4	12	3
Bank Street	05	14	14	5	5	1	5	5	4	6	8	3
Mathemagenic	06	7	5	1		1					3	3
Oregon	07	20	16	5	6	1	4	5	4	5	10	3
Behavior Analysis	08	13	13	3	4	1	6	5	6	6	8	3
Cognitive Curriculum	09	10	7	3		1	6	5	6	3	6	3
Fla. Parent Education	10	11	11	3	6	1	5		5	4	9	3
EDC Open Education	11	10	10	3	5	1	5	5	5	3	8	3
Indiv. Early Learning	12	7	7			1	3		2		3	3
Interdependent Learn.	13	3	2	2	2	1				1	2	2
Language Develop.	14	5	4	1		1	3		3	3	5	3
Parent Implemented	15	14	12	1								
Hampton Non-Graded Mod.	17	6	4									3
Cult. Ling. Approach	18	5	3								2	3
Parent Supported	19	2	2									2
Respon. Environs.	20	1	1									1
Home-School Partnership	21	2	2							1	1	2
Cal. Process Model	22	6	****							4	5	
New School Approach	23	4	4								2	3
Afram Par. Implemen.	24	9	10****									
Cult. Demo. Learning	26	1	1									1
Role Trade Model	27	1	1									1
TOTALS		199	176	39	38	12	48	36	47	52	100	51

*Each of the major national evaluation reports includes more than one analysis. Summarized here are the distributions of sites included in the broadest analyses from each of the reports. Specifically these are: Emrick - Cohort I, one- or two-year effect study; Soar - Cohorts I and II, classroom observation study; Stallings 1971-72, classroom observation study; Abt I - Cohort III, Kindergarten one-year effects study; Stallings, 1972-73 - Cohort I third grade and Cohort III first grade classroom observation; Abt II - Cohort III kindergarten study; Abt III - Cohort II third grade; Abt IV - Cohorts II and III third grade; and RMC overall cost analysis.

**These totals are taken from SRI, Follow Through Evaluation Data Bank File Description All Years Beginning 1972-73, Part 2, Appendix A. They sum to 199 because some projects are associated with more than one sponsor. There were actually only 182 different projects.

***The total number of different projects in Follow Through as of July 1976 was 164. The total given here sums to 176 because some projects had multiple affiliations. Source: Division of Follow Through, Bureau of Elementary and Secondary Education, U.S. Office of Education (Fred Bresnick, November 1976).

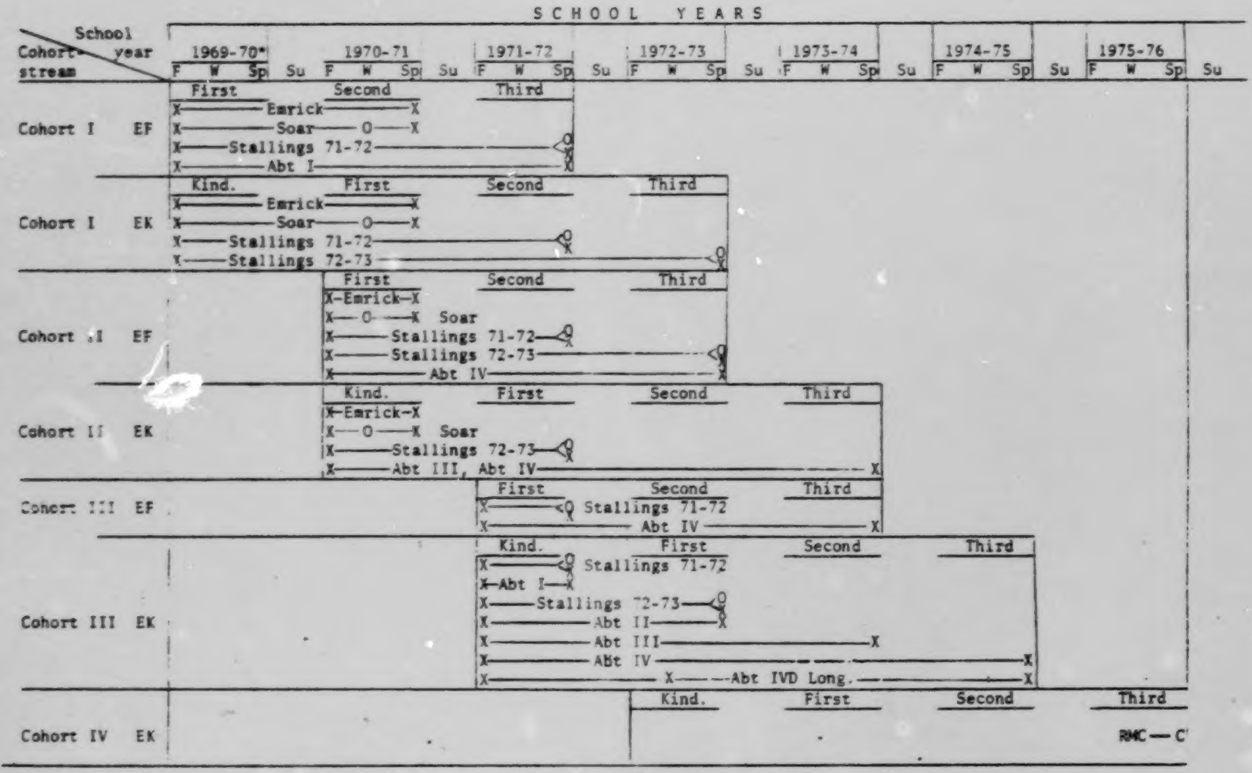
****The Afram and the California Process Models were not included in Follow Through after 1974-75. Hence the projects associated with these sponsors are now categorized together as unsponsored--they differ from self-sponsored models in that they once were sponsored.

The cohorts and cohort streams treated in the different major studies are represented in Figure 4. As can be seen in this figure, the classroom observation studies by Soar and Stallings generally were carried out on cohorts I and II. Not shown here but as is indicated in Table 5, the classroom observation studies generally dealt with a relatively small sample of Follow Through projects--38 in the Soar study, 12 in the 1971-72 Stallings report, and 36 in the 1972-73 Stallings study.

Figure 4 also shows that the entering-kingergarten stream of cohort III has weighed most heavily in evaluation efforts. More studies have included this cohort-stream than any other. Note also that the vast majority of studies in these major national reports have employed a pre-post design analyzing data from children's entry into Follow Through and from some subsequent time point. The figure also shows that the RMC cost study employed data gathered in spring 1976--after the first three evaluation cohorts had already graduated from Follow Through.

The evaluation's emphasis on a limited number of sponsors, projects and cohorts results partially from necessity. It would have been extremely expensive to include every cohort in all projects associated with every sponsor. Also, the focus on the original twelve sponsors resulted from the [illegible]70 decision to exclude sponsor start-up years from evaluation. Since sponsors #17 through #27 did not join the program until 1969 or later their first year sites were considered "implementation year projects" and so not eligible for evaluation under the start-up year policy.

Figure 4: Time and Cohort of Data Included in Major National Reports*



Key:
0 - Observation data collected
X - Test data collected
C - Cost data collected

Notes: Not indicated in this figure are teacher questionnaire and parent interview data, which will be discussed in chapter 4.

1. Emrick - Although not indicated in the above chart the Emrick report also included classroom observation data from spring 1971 for a subset of cohort I and II sites covered in the report. Also not indicated above but included in the Emrick report are cost data "focused mainly on the school year 1970-71" (p. 55) collected over a period "extending into spring 1972" (p. 54).
2. Soar - Preliminary reports by Soar's group also included classroom observation data gathered in the winters of 1969 and 1970. The bulk of the Soar report focused on the classroom observation data rather than on test data.
3. Stallings 71-72 - The bulk of the first Stallings report dealt with observation data gathered in spring 1972 rather than on test data.
4. Abt I - The Abt I analysis of cohort I data was for a relatively small set of projects - 25 projects associated with six different sponsors (p. VII-135).
5. Stallings 72-73 - The bulk of the second Stallings study also focused on classroom observation data.
6. Abt II - Monograph appended to the Abt II report made some use of spring kindergarten data for cohort III-EK and briefly described the availability of cohort IV data.
7. Abt III - The Abt III report focused on the cohort II data, and gave relatively little attention to the cohort III data.
8. Abt IV - The bulk of the Abt IV report dealt with pre-post analyses of data for four cohort streams--CII EF, CII EK, CIII EF and CIII EK--but one part, Abt IVD, dealt with data for cohort III EK at three time points
9. RMC - The RMC cost data were gathered from March 15 to May 14, 1976, by which period all but the cohort IV EK children had already graduated from the program.

* Fall-winter 1969 testing - "Because of the scope and magnitude of the fall 1969 effort, and because of delays in negotiating the content of the test battery, completion of the cohort I baseline testing often occurred as late as December 1969, nearly half-way into the school year" (Emrick, Interim Evaluation, February 1973, p. 49).

The focus on fewer sponsors and more sites per sponsor in the Stallings 1972-73 report and the first two Abt studies was part of the planned variation ethos in the sampling design set out by Jane David in 1971. In explaining the "attempt to minimize differences among sites with different sponsors, and when possible to include at least two similar sites per sponsor in cohorts I and II and at least five sites per sponsor in cohort III," David wrote:

> We included as many diverse sites as possible in order to decrease bias and increase precision in the estimation of effects. Since Follow Through is not a true experiment, randomization did not occur at any level of assignment of treatment. Thus it is difficult to know if groups being compared differ only in treatment. Since some child and site characteristics were measured we know that confounding exists and we know to what extent it exists on the measured variables. But without randomization we do not know whether confounding exists on the measured variables. Thus all estimates of treatment effects are biased in some unknown way. If a large number of different projects are included for a given sponsor, then we can be fairly confident that the remaining biases are not strong. Additionally, the more projects there are per sponsor, the more precise is the estimate of treatment effects because of increased degrees of freedom in the error term.[8]

This rationale contributed directly to the focus in Abt I and II on a set of only ten sponsors. Only for these sponsors with data for several sites per sponsor would it be possible, according to David's rationale, to avoid complete confounding of sponsor-treatment effects estimates with site characteristics. In plainer language this meant simply that it was best to try out each model in four or five sites in order to make sure that judgments on the effectiveness of a model did not depend solely on what happened in one particular place. McDaniels later summarized this policy: "In 1971, sponsors with fewer than three sites or an unrepresentative distribution of sites were eliminated from the FT evaluation."[9]

---

[8] Jane David, Final Report on Sample Selection (Cambridge, MA: Huron Institute, June 30, 1973), pp. 3-4. The rationale is interesting in several respects. First of all, it rests on implicit assumptions about independence of sites matched to particular sponsors. Second is the implicit but clear assumption that the main goal of the evaluation was to estimate sponsor treatment effects.

This reasoning also led to the focus on cohort III as the main evaluation cohort. Because of unsystematic testing in prior years, cohort III provided the best opportunity to estimate sponsor treatment effects across several different sites per sponsor. According to David's Final Report on Sample Selection, 182 different sites had been included in the Follow Through program, but of these only 84 cohort III sites would be tested at the end of third grade--66 in the entering kindergarten stream and eighteen in the entering first stream. Most of the cohort III sites not included in the sampling plan were excluded because of insufficient baseline data, or because they had dropped out of the program--but more than a dozen sites were not included simply because there were "not enough comparable...sites for the sponsor."[1]

The Abt III and IV and the RMC reports reflect an implicit shift away from this sort of thinking. Both these studies have opted for breadth of coverage, including data for as many sponsors as possible, even though data for some sponsors are drawn from only a single site, and hence some sponsor "effects" or "costs" are completely confounded with the effects and costs of the particular sites with which they work.

---

[9] Garry McDaniels, "Evaluation Problems in Follow Through," in Rivlin and Timpane (eds.), Planned Variation in Follow Through, p. 54.

[1] J. David, Final Report on Sample Selection, attachment 8.

Focus of Inquiry

Foregoing pages outlined the scope of the national Follow Through evaluation in terms of possible evaluation criteria, program components, sponsors, projects and cohorts included in the various national evaluation studies. These dimensions are useful to delineate the general scope of the national evaluation effort, but they are highly concrete and leave us short of a more important dimension to the inquiry, namely, the kinds of questions addressed in the national evaluation studies.

The domain of issues that might be addressed in a program evaluation can be roughly broken down into four types of questions:

--What is the program?

--Does the program make any difference in the lives of people it was designed to serve?

--How does the program work to make a difference?

--How much does the program cost?

These four areas of inquiry correspond roughly to what Edward Suchman calls effort, performance, process and efficiency evaluation.[2] The categories obviously overlap and, in practice, are highly interdependent. Nevertheless, these distinctions can help locate the focus of inquiry of the evaluation effort.

What is the Follow Through Program? In 1971 Richard Fairley began his testimony on Follow Through before the U.S. Senate Committee on Appropriations, "The Follow Through program is designed to capitalize...."

---

[2] E.A. Suchman, Evaluation Research (NY: Russell Sage, 1967), pp. 61-7. Suchman also mentions a fifth category--adequacy of performance, but for present purposes this will not be considered as a separate category since performance standards for Follow Through were never established.

Before he could finish, Senator Warren Magnuson interjected, "I don't care what it is designed for..I want to know what you do. I know what it is designed for."[3] After the issue of purpose, the natural next question is the descriptive one. What is the program? As Senator Magnuson inquired, what does the program do?

Despite its obviousness the descriptive question--especially as it applies to the overall Follow Through program--has received remarkably little attention. The bulk of the national evaluation has focused on the planned variation aspects of Follow Through and not on other program components like comprehensive services.[4] Within the planned variation domain, there also seems to have been relatively little attention paid to descriptive questions. Most descriptions of the planned variations of Follow Through rely heavily upon Follow Through sponsors' descriptions of intent and only rarely are based upon empirical descriptions of how sponsors' curriculum models are actually translated into classroom practice.

Among national evaluation studies, the most relevant to any empirical description of Follow Through planned variations are the classroom observation studies conducted by Robert Soar of the University of Florida and Jane Stallings of the Stanford Research Institute. Both these observation studies were remarkable in surmounting tremendous procedural and measurement issues, and they are described in more detail in chapter III. Nevertheless, in comparison to the scope of the planned variations, the classroom observation studies have been relatively small, each encompassing fewer than 40 projects. Also, both the Soar report and the second Stallings study were heavily oriented toward investigating relationships between variables and developed highly abstract variables--in Stallings' work, standardized implementation scores and in Soar's case, factor scores--which do not afford very clear pictures of what the Follow Through planned variations actually are in terms of classroom practice. So among the 7000 pages of the nine major reports only around 100 pages of the 1971-72 Stallings study give direct description of what goes on in Follow Through classrooms.[5]

---

[3] U.S. Congress, Senate, Senate Hearings Before the Committee on Appropriations Office of Education, Special Institutions and Related Agencies Appropriations, Fiscal Year 1972 (Washington, D.C.: U.S. GPO, 1971), p. 935.

[4] In its 1975 review of Follow Through, the GAO criticized OE's inattention to Follow Through non-instructional services. The GAO criticized OE monitoring of these services as "ineffective," and, among other things, recommended that the Secretary of HEW direct the Office of Education to "establish to the extent practicable, performance standards for each service component and require LEA's to evaluate their progress toward meeting these standards." Comptroller General, Follow Through, pp. 27, 38. In apparent response, the Follow Through program office commissioned Boon, Young and Associates of New York to conduct a survey of Follow Through comprehensive services. The survey report was originally scheduled for completion in June 1977, but has been delayed until October 1977.

Does the program work? Does the Follow Through program work; does it make any difference for the children it was intended to serve? This is the sort of question which has clearly predominated in the Follow Through evaluation effort. The form of this question stated in the 1972 RFP for analysis of evaluation data was:

> -- Is there any overall difference in the impact of Follow Through and comparison schools? That is, does the Follow Through Program in the aggregate have a greater impact on parents and children than non-Follow Through schooling?
>
> -- Are there educationally significant differences in the impact of the various educational strategies being tested in Follow Through?[6]

---

[5] Stallings, 1971-72, pp. 79-187. Another Follow Through evaluation study, J.C. Grannis, The Columbia Classroom Environment Project...Fifth Progress Report (New York: Columbia University Institute for Pedagogical Studies, May 1972), also provides direct information on what goes on in Follow Through classrooms. These descriptive studies will be treated more fully in later chapters.

[6] OE, RFP, Follow Through, Draft, March 13, 1972.

Since 1972, attention has funnelled even more narrowly to focus on the second of these questions. The overall Follow Through impact question was dropped from the 1974 evaluation RFP, and the issue of program impact on parents has eroded away as questions asked of parents shifted from year to year and analyses of parent interview data varied.

So by 1976, the national Follow Through evaluation has narrowed to focus on whether the various educational strategies being tested in Follow Through have educationally significant effects on participating children. Since the 1972 reorganization limited the evaluation to a standardized test battery, the domain of educational significance has been limited to the MAT, the Ravens, the Coopersmith and the IARS. The magnitude of educational or "material" significance has been set at one-quarter standard deviation of each outcome variable.[7] Focus on the impact issue and on these dimensions of educational significance has caused considerable consternation as mentioned in chapter I. In chapter III we recount in more detail why this focus evolved, so for now let us turn to the third area of inquiry.

How does the program work? Since virtually the beginning of the program, Follow Through evaluators have recognized the importance of the process issue. Recall that even before the planned variation evaluations, Kurfeerst and Stephens wrote:

---

[7] To be more precise the determination of "educational" or "material" significance has been based on the pooled within-site variance of each outcome variable (see Abt III, p. A-74). I have not been able to track down the exact origin of the 1/4-standard-deviation-equals-material-significance policy, but the definition seems to appear first in the Abt II report (p. V-5).

> Answering the question, "What changes took place in the Follow Through population over the period of the program" is essentially meaningless in terms of the evaluative function if one cannot also answer the question, "What actually occurred to produce the changes?"[8]

The importance of the process question was reflected in different form in the Emrick report in 1973:

> It was clear early in the planning and preliminary evaluation activities that for FT to be evaluated as a total program, more must be measured than the participating child's academic progress. Furthermore, since evaluation interest would be focused on identification of "components" or "successful" programs, attention would need to be given to evaluating the process as well as the outcome.[9]

The significance of the process issue was stressed once again in Abt's 1974 proposal. After "establishing the differences,...[or] finding out how outcomes of the various strategies differ," wrote the Abt investigators, "the appropriate next step is to be as certain as possible that the findings represent something more than a quirk of data or circumstance and to discover why they came out as they did."[1]

The Follow Through evaluation effort has edged over into the process portion of inquiry in two ways. First, the classroom observation studies by Soar and by Stallings have attempted to link differences in classroom processes to differences in test scores. Soar did this mainly through correlational analysis and Stallings through partial correlation and multiple regression analyses. Also, the Emrick and Abt evaluations indirectly addressed the process issue in analyses designed to estimate effects over different time periods (for example, one-year effects or two-year effects),

---

[8] Kurfeerst and Stephens, pp. 69-70.

[9] Emrick, Interim Evaluation, p. 33.

[1] Abt, "Proposal," 1974, p. 7.

at different sites, or for different kinds of children (for example, children in urban schools versus those in rural schools, or for students with different kinds of preschool experience).

Other kinds of process questions have received relatively little attention in the nine major national evaluation studies. For example, the issue of how sponsors' curriculum models are translated into reality in the classroom is touched upon only in the 1971-72 Stallings report, appendix K, "Reports of Staff Training by Sponsors."[2] The three classroom observation studies also deal with the relationships between classroom processes and children's test scores. Other process issues--like the relationship between model implementation and model impact on student learning, the connection between program effects and parent participation, and a host of questions on the inter-relationships between effects, classroom practices, and non-classroom components of Follow Through--have occasionally been raised but rarely directly addressed in any of the major national studies.

How much does the program cost? Like other questions, the issue of cost has often been raised but has been little answered. Even before the planned variation phase of Follow Through, the Pittsburgh evaluators called for assessment of the costs of Follow Through. Among major national studies prior to the RMC report only the Emrick volume reports cost data.

---

[2] Stallings 1971-72, pp. 435-447. A generalized description of Follow Through program implementation is also given in D. Judd, Project Follow Through: A Description of Sponsor Implementation Processes (Portland: Nero & Associates, April, 1975). See bibliography.

Estimates of Follow Through per pupil costs have ranged from $400 to $1200 with the average around $800. These estimates as well as the RMC cost study will be summarized in chapter V. For now, suffice it to note that like most other aspects of the national evaluation, the cost study has been the object of much argument. More than a year before the RMC study was finished, at their 1975 annual conference, Follow Through sponsors resolved:

> Whereas Sponsors have had nominal control of the total amount of dollars available to local sites, whereas sponsors have had nominal control of program implementation in local sites, and whereas sponsors have not been given the opportunity to advise on this proposed work; it is:
>
> Resolved that while the RMC cost study will produce important information for Title I and Follow Through, it will not reflect sponsor implementation costs or the actual costs of Follow Through.[3]

---

[3] High/Scope. Report of the Fifth Annual Follow Through Sponsor Conference. Ypsilanti, MI: High/Scope Foundation, October 1975, p. 3.

* * * * * * * * * * * * * * *

This survey of the scope of Follow Through national evaluation studies shows several patterns. First, the number of sponsors included in the nine major studies has gradually grown--from 10-12 in early studies to 17 in the Abt IV report and 20 in the RMC cost study. So in terms of breadth of sponsors treated, the evaluation effort has steadily expanded.

A second pattern, however, is that the sampling of different projects associated with sponsors and of the different cohort-streams has been highly varied. The bulk of the classroom observation studies, for example, were concentrated on cohorts I and II and on data gathered in the winter of 1971, the spring of 1972 and the spring of 1973. In contrast the bulk of the Abt outcome evaluations have focused on cohorts II and III, on data gathered in the spring of 1974, and the spring of 1975. Furthermore, the RMC cost study relies upon information gathered in spring of 1976, a year after the third experimental cohort had already graduated from Follow Through. The diversity of the cohorts, projects and times sampled in the different national studies makes it clear that no mechanical aggregation of findings from them will be possible.

Perhaps the most interesting pattern apparent in this review is that the nature of inquiry has undergone a sort of funnel vision, with the sort of questions asked gradually narrowing to focus in the Abt IV report on whether sponsors have importantly different effects on children, as measured by the four tests in the third grade outcome battery. Other questions--having to do with classroom processes, overall Follow Through program effects, comprehensive service delivery, and program effects on parents and other program participants--have gradually been dropped from the national evaluation.

The causes of the funnel vision have been varied. The narrowing was partly due to political pressures on the evaluation to produce results. Also it was clearly related to the shift in responsibility for the evaluation--from the Follow Through program office to the Office of Planning, Budgeting and Evaluation. More fundamentally, however, the narrowing view has been a product of the way in which people think about evaluation. This thinking reflects the predominant paradigm in evaluation research--the model of the treatment-control experiment. In chapter III we examine the way in which this paradigm has been used in the Follow Through evaluation and the way in which the various features of the paradigm have contributed to the evaluation of Follow Through that we now have.

III. ISSUES IN THE ANALYSIS OF STUDENT OUTCOME DATA

> There is almost universal agreement among evaluation researchers that the randomized controlled experiment is the ideal model for evaluating the effectiveness of a public policy.
>
> \- Peter Rossi and Sonia Wright[1]

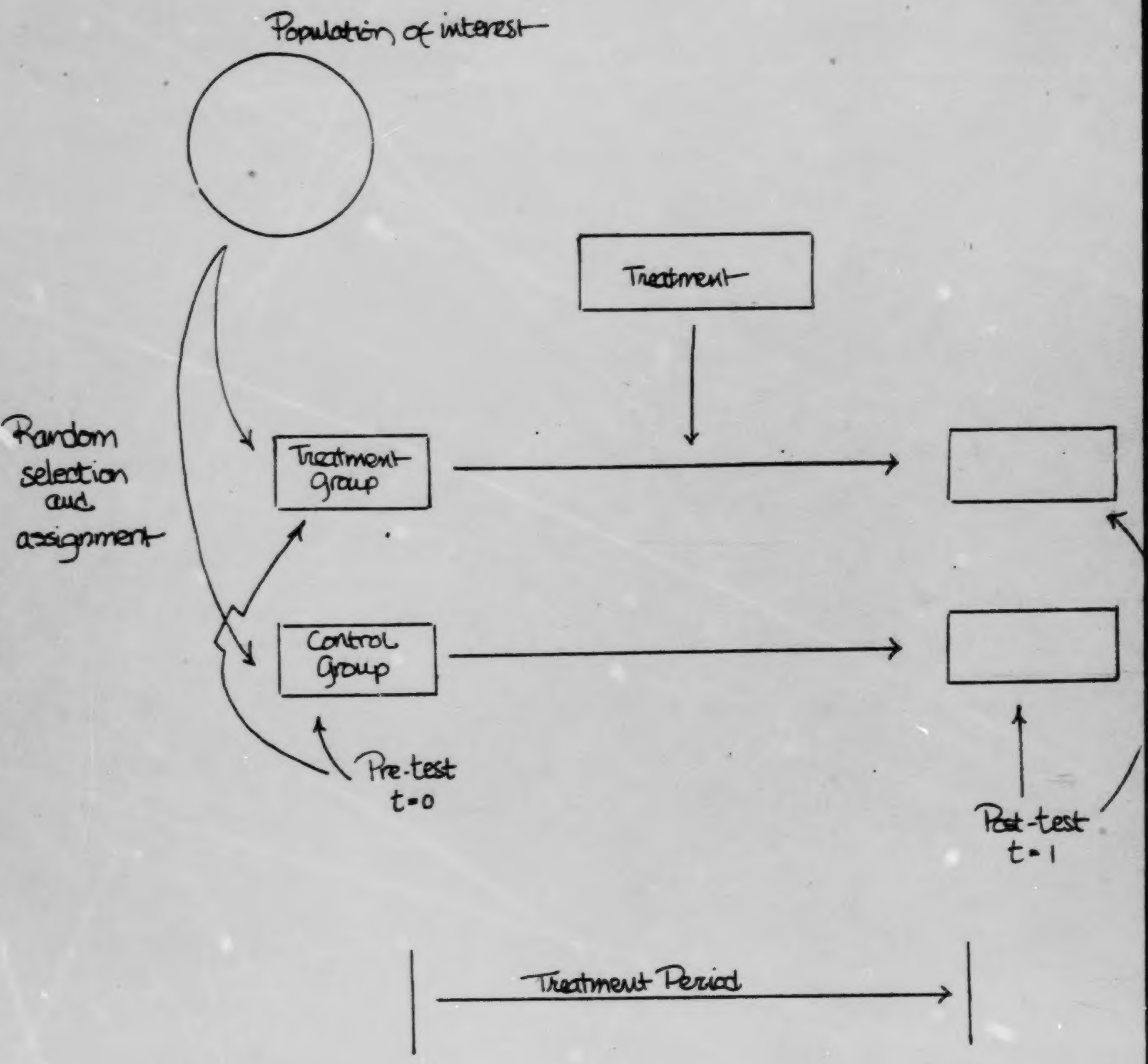
How can an investigator determine the effects of a program? This is the question which faces the evaluator wanting to know whether or not a program works. There are a variety of methods that could be used, but the ideal prevailing recently among evaluators is the treatment-control experimental paradigm.[2] The idea, as illustrated in Figure 5, is quite simple. The investigator begins with a population of interest and from it selects two equivalent groups. One group is assigned to the treatment condition and becomes the treatment group. The other receives no treatment and so becomes the control group. Before starting the experiment, the two groups are observed or pretested to make sure they are roughly the same on all important measures. Then the treatment group is given the treatment for a specified time and the control group is not. At the end of the treatment period, each group is tested again, and the differences in the post-test scores for the two groups provide an estimate of the treatment effect.

---

[1] Peter Rossi and Sonia Wright, "Evaluation Research. An Assessment of Theory, Practice and Politics," Evaluation Quarterly, Vol. I, no. 1, February 1977, pp. 5-47.

[2] The paradigm described here is a simplified version of the first of Campbell and Stanley's true experimental designs, that is, the "pretest-posttest control group design." D.T. Campbell and J.C. Stanley, "Experimental and Quasi-Experimental Designs for Research" in N.L. Gage (ed.), Handbook of Research on Teaching (Chicago: Rand McNally, 1963).

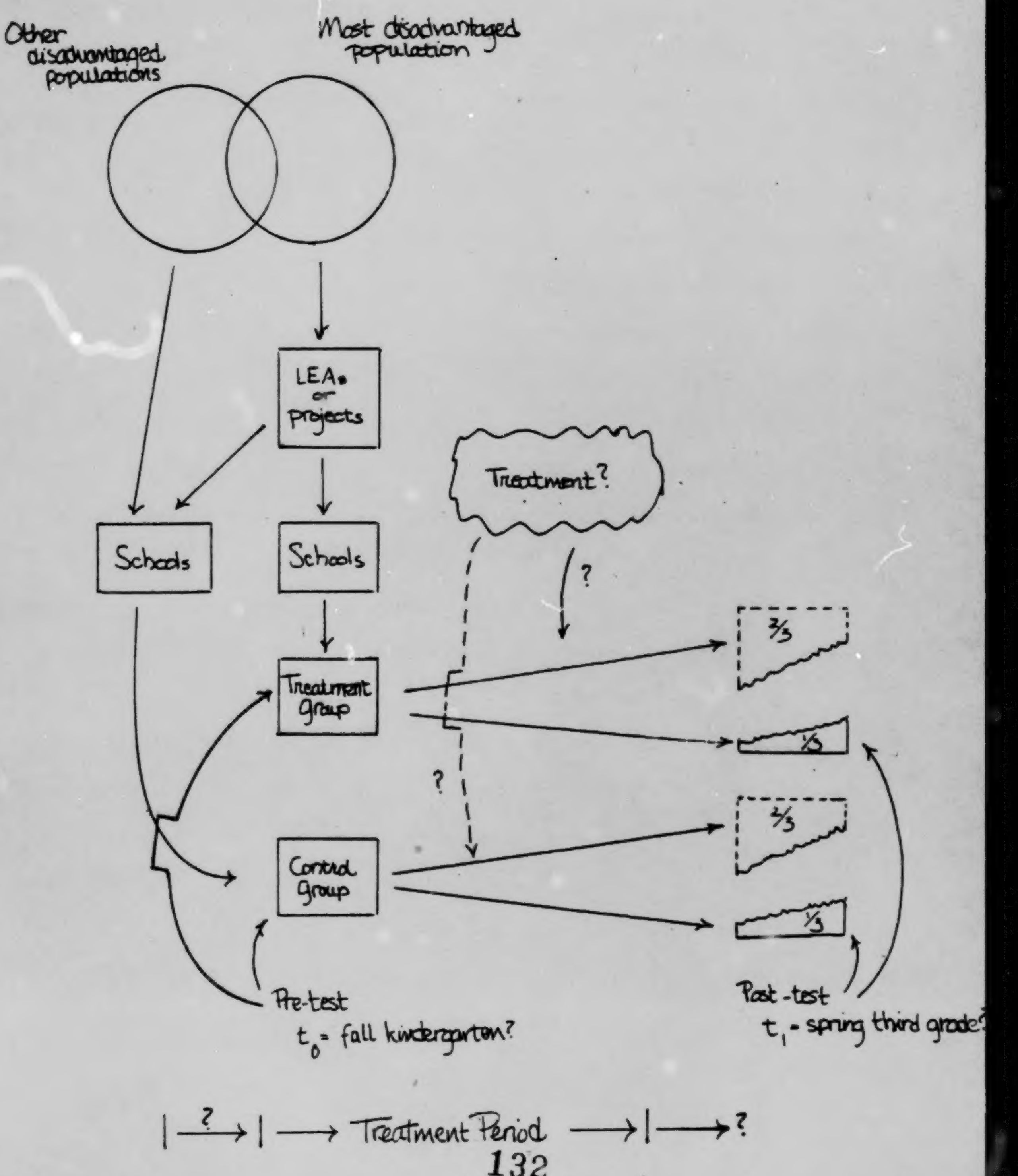
Figure 5: The Randomized Controlled Experiment Ideal



Nevertheless, the Follow Through planned variation experiment falls short of this experimental ideal in nearly every aspect, as is shown in Figure 6. First of all, the Follow Through treatment groups and the control groups (Non-Follow Through, or NFT) do not seem to come from the same population and so are generally not initially equivalent at the time of pretest. More fundamentally, it is unclear exactly what comprises the Follow Through and Non-Follow Through groups, that is, what units in the groups should be compared. Should the Follow Through evaluation compare results for individual children, for individual classes, individual schools, sponsors or what? Also, the exact nature of the Follow Through "treatment" and how to verify its administration are both unclear. What is a Follow Through treatment and how does the evaluator know whether it has been given? Moreover, how does the evaluator know whether the control group has not received the treatment? How can evaluators tell whether or not the Follow Through treatment may have diffused or spilled over to the control group?

A related problem is the appropriate time period between pre-test and post-test. Although entry into kindergarten or first grade has generally been assumed to be the appropriate pre-test time and the end of third grade the appropriate post-test time, these boundaries for the treatment period have both been challenged. Furthermore, there is no clear-cut standard of comparison to be used in determining the presence or absence of program impact at post-test time. What difference on what tests or outcome measures should be used to determine whether there are any "effects" of Follow Through? And still another major problem: Given that not all children in Follow Through and Non-Follow Through groups who were present for pre-tests are

Figure 6: Quasi-experimental Reality in Follow Through


Other disadvantaged populations
Most disadvantaged population
LEAs or projects
Schools
Schools
Treatment?
?
Treatment Group
Control Group
2/3
1/3
2/3
1/3
?
Pre-test
t_0 = fall kindergarten?
Post-test
t_1 = spring third grade?
? → | → Treatment Period → | → ?

still in the program when it comes to post-test time, what should be done? How should this problem of attrition be handled?

Given all of these problems, the question of how pre-test and post-test scores are to be analyzed becomes far more difficult than if the "planned variation experiment" had more nearly approximated the ideal. One of the most common statistical techniques for analyzing data from quasi-experiments like Follow Through, where treatment and comparison groups are not initially equivalent, is the analysis of covariance, or ANCOVA. ANCOVA is a statistical technique which can help the evaluator control for such initial differences. As such it has been widely used in the Follow Through evaluation, both by SRI and by Abt. But the question still remains, how good is it? In the words of a recent article on this technique, is the ANCOVA an angel of salvation, or a temptress and deluder?

In this chapter, we take up each of these analytical problems:

-- selection and assignment
-- units and levels of analysis
-- specification and verification of treatment conditions
-- period of treatment
-- standards of comparison and outcome measures
-- attrition
-- ANCOVA

While these analytical issues are separated for the sake of discussion, they are in actuality deeply interwoven in the complexity of the Follow Through planned variation experiment. Differential selection and assignment to treatment and control groups necessitate some sort of analysis like ANCOVA to compensate for initial differences between treatment and control groups. Follow Through treatments are not very well-defined, but different ones would seem to require different sorts of outcome measures. Different sponsors' approaches, it has been claimed, implicitly require treatment periods of varying length. But different treatment periods will produce different levels of attrition, and differential attrition, it has been argued, may necessitate different units-of-analysis. Yet different units, it has been reasoned, required different ANCOVA's. And so it goes. These issues have been interargued in the Follow Through evaluation to such an extent that consideration of any one quickly entangles us in almost all of the others.

Yet for the sake of analysis, the issues must be treated one at a time. As each is considered, we recount how it has troubled the evaluation and how it has been handled in the major Follow Through evaluation reports. We begin with the issue of selection and assignment which is the heaviest strand in the web of complexity surrounding the Follow Through evaluation analyses.

Selection and Assignment

> Random assignment of study subjects to experimental or control groups, is the essential feature of true experiments.
>
> \- Henry Riecken and Robert Boruch[3]

---

[3] H. Riecken and R. Boruch (eds.), Social Experimentation. A Method for Planning and Evaluating Social Intervention (New York: Academic Press, 1974), p. 4.

The crux of the experimental paradigm just described is that treatment and control groups start out the same. As Donald Campbell and Julian Stanley observed in their famous paper on the design of experimental research, randomization "is a less than perfect way of assuring the initial equivalence of such groups. It is, nonetheless, the only way of doing so, and the essential way."[4] The problem is that if individuals are not randomly assigned to treatment and control groups, then the two groups may not be essentially the same at the start of the experiment. Hence, it becomes doubtful whether to attribute later differences to treatment effects or to initial differences or to some combination of the two. To explore this doubt, we need to look at several aspects of the selection and assignment problem. Specifically, the following paragraphs discuss: the lure of randomization; the selection of treatments, the selection and assignment of projects to treatments; the selection of schools and children; the selection of evaluation sample; and the selection of control groups.

The Lure of Randomization is that it can help to make treatment-control experiment simple. If subjects are randomly selected from a population of interest then one can more safely infer results of treatment-control experiments to the population from which experimental subjects are drawn. If subjects are randomly assigned to treatment and control groups, this helps to insure that these groups start out essentially equivalent. This makes it safer to infer that differences between treatment and control groups that show up at the end of the treatment period are in fact attributable to the treatment.

---

[4] D.T. Campbell and J.C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963). Randomization is relevant to issues of external as well as internal validity of results. External validity is vital to the generalizability of research findings, but discussion here is limited to the problem of internal validity, that is, the validity of the findings in the context from which they are derived.

Despite its lure, randomization has been used virtually nowhere in the Follow Through program. So Follow Through evaluation studies have generally been able to do little about nonrandom assignment save to note it as a problem, and to cite it as a cause for using special analytical techniques like ANCOVA to adjust for initial nonequivalence of treatment and control groups.[5] One treatment of the randomization issue in the Follow Through context was offered, surprisingly enough, in the Government Accounting Office's 1975 review of Follow Through. The GAO reviewers wrote:

> Although the evaluation was substantially revised during the fourth year of the experiment, problems in both initial design and implementation of the experiment will, in the final analysis, limit OE's ability to reach statistically reliable conclusions as to which sponsor approaches are successful for teaching young disadvantaged children. These problems, which we believe cannot practicably be overcome include
>
> -- lack of random assignment of LEAs [or projects] to sponsors and
>
> -- lack of comparable control groups.[6]

The GAO critique implied that random assignment would have been a useful technique to employ in the evaluation. The suggestion did not discuss the practical and political problems involved in trying to use random assignment with a program as large as Follow Through. Also, it was somewhat ironic because it ignored the fact that random assignment had already been attempted in one small corner of the Follow Through evaluation.

---

[5] For example, in Emrick, pp. xxxvii, 11-12, Abt I, p. 11, and Abt IVA, p. 13.

[6] Comptroller General, Follow Through, p. 6.

One of the supplementary studies commissioned by Garry McDaniels in 1972 was what came to be called the Follow Through summer study. The study was designed to investigate "the impact of exposure to schooling on achievement growth of children" and was conducted by Jane David of the Huron Institute.[7] In outlining how to pursue her investigation, David wrote, "The ideal study of these issues would be an experiment which involved random assignment of children, perhaps blocked on social class, to school and no school."[8] One part of David's research actually attempted the "ideal" of randomization. Her plan called for the random selection of children from a list of eligible and interested students for assignment to a summer program treatment group. The students not selected for program participation would form the control group.

David's plan was to apply the random assignment procedure in eleven Follow Through projects. One project initially dropped out of the summer program. Then, David reported, "serious problems in following the procedures arose in four [more] of them." In three of the sites, summer programs had been scheduled which would interfere with the status of the "control" group. In the fourth project there were not enough eligible children to divide between treatment and control groups. David continued,

---

[7] Jane David, Summer Study: A Two-Part Investigation of the Impact of Exposure to Schooling on Achievement Growth. Ed. D. thesis, Harvard University, 1974. The idea for a Follow Through summer effects study originated before either McDaniels or David became involved in Follow Through; from whom I do not know.

[8] David, Summer Study, p. 22.

> This left six projects still included in the Summer Project Study. Of these six, two projects had only six children in the control group who had received all tests. Consequently these were not included in the analysis. Of the four remaining for which the data were analyzed, each has its own peculiarities. (p. 84)

In one of the sites, peculiarities were sufficient "to raise questions about the adequacy of the randomization procedure. Consequently, it seems impossible to draw any conclusions from these data" (pp. 99-100). Another problem cropping up in the randomized summer study was whether children randomly assigned to a summer program actually attended. Parents of a number of children assigned to the summer school treatment group "claimed" that their children did not attend a summer program (p. 85). After investigating this discrepancy, David identified a group of subjects for whom the randomized treatment-control conditions seemed to hold. And for the three projects where this seemed to be the case, she concluded "the findings indicate that the summer programs do have a short term impact (as do most programs)" (p. 104).

Without detailing more of David's study, the example makes it clear that randomization did not prove to be the panacea that one might hope. Despite randomization, David's study still faced many of the problems confronted in the larger Follow Through evaluation, including attrition and specification, and verification of the treatment. These problems arose even though the summer program in David's study was midget-sized compared to Follow Through and lasted only three months instead of four years.

In surmounting the obstacles of attrition, treatment verification and other confounding problems in order to answer the program impact question, David had to dispense with rival hypotheses which might plausibly explain the apparent program impact she found. And this is precisely what an investigation must do in the nonrandomized case as well. As Campbell and Ross have written:

> An absence of randomization may in some specific way plausibly explain the obtained results. But unless one can specify such a hypothesis, and the direction of its effects, it should not be regarded as invalidating.[9]

The major problem is that the absence of randomization opens a whole new realm--initial nonequivalence of treatment and control groups--from which can spring hypotheses to threaten the program effect inference. So in the absence of random assignment, it is important to consider how subjects were in fact selected and assigned to treatments. First, however, we pause to consider how the Follow Through "treatments" themselves were selected.

Selection of Program Treatment. If an experimenter wants to investigate whether any of a whole range of treatments have effects, he or she may randomly select a sample of treatments to try out, using what is known as a "random effects" model of analysis.[1] This approach is relatively rarely used in practice and was not used in Follow Through because interest was not in some wider set of treatments but in the specific set of conditions prescribed in Follow Through sponsors' models. So from a technical point of view, a fixed effects approach is clearly appropriate to Follow Through. Nevertheless, it is of substantive interest that the different sponsor model approaches used in Follow Through were selected for different sorts of reasons.

---

[9] D.T. Campbell and H.L. Ross, "The Connecticut Crackdown on Speeding: Time Series Data in Quasi-Experimental Analysis." In E.R. Tufte (ed.), The Quantitative Analysis of Social Problems (Reading, MA: Addison-Wesley, 1970), p. 123.

[1] G.W. Snedecor and W.G. Cochran, Statistical Methods (Ames, Iowa: Iowa State University Press, 6th edition, 1967), p. 364.

The original set of Follow Through sponsors was selected after a canvas of "persons who had developed new approaches for working with young children or who had access to information about new developments." Following review of potential sponsors "eighteen groups which appeared to have relatively well-developed ideas were invited to present proposals to Follow Through communities. Sixteen accepted the invitation and fourteen were chosen by the first set of communities in 1968."[2]

Six additional sponsors entered the program in 1969-70, at least in part to give "opportunity for representation by three different groups--State education agency, minority colleges, and profit-making company--not included in the first set of sponsors."[3] The last four sponsors were added to the program, at least partly, to remedy "under-representation of minority institutions."[4]

---

[2] Egbert, "Follow Through," pp. 10-11. The number fourteen includes "parent implementation as a type." One of the original fourteen--the DARCEE model sponsored by George Peabody College--dropped out of Follow Through in 1970.

[3] Egbert, "Follow Through," p. 13. For a fuller account of the selection of the first two sets of Follow Through see Elmore, Follow Through: Decision-making, pp. 197-210.

[4] Egbert, "Follow Through," p. 30.

Selection and Assignment of Projects. Like Follow Through sponsors, Follow Through projects were selected in several stages. The first set of Follow Through projects was chosen in 1967. Some 500 communities were originally nominated for Follow Through by state education agencies, and during the first week in May 1967, some 100 communities were invited to submit proposals to participate in the new program. Decisions on whom to invite were made by OE officials in consultation with OEO representatives. A selection panel convened by OE reviewed the 87 proposals submitted and "On June 15, Harold Howe II, Commissioner of Education approved the 30 pilot centers."[5] A second set of ten pilot sites was selected in late summer 1967 to join the program during the second semester of the 1967-68 school year in response "to pressure from big city mayors and congressional delegations."[6]

In addition to direct political considerations, selection of the 40 pilot sites in 1967 was guided by general provisions of the Economic Opportunity Act amendments then before Congress, including provisions for urban-rural distribution of funds, limitations on distribution of funds among the states, and plans for participation of parents in the program.[7] To these general requirements, another special requirement was added to the selection process in 1968--that new communities would be required to select one of the original fourteen sponsor approaches. Otherwise the 1968 stage in the project selection seems to have resembled the first. Some 200 school districts were nominated by state education and economic opportunity agencies, and from them 52 were selected to join the new planned variation phase of Follow Through.

---

[5] Egbert, "Follow Through," p. 5.

[6] Egbert, "Planned Variation in Follow Through," pp. 2-4.

[7] Egbert, "Follow Through," p. 8.

The "assignment" of projects to program treatments took place at the two sets of meetings in Kansas City in February 1968--the first to match pilot sites with sponsors and the second to match the new projects with sponsors. Accounts differ as to whether pilot sites were required to affiliate with sponsors or not, but most did so. The misunderstanding later resulted in hard feelings because officials in at least one pilot project felt they were coerced into choosing a sponsor.[8]

Grounds for match-ups between sponsors and projects ranged widely; from chance common background of project and sponsor personnel, to results of a formal election by parent representatives, to advice of Follow Through general consultants.[9] In general, discretion for the pairing of projects with sponsors--in the experimental idiom, assignment of treatment groups to treatments--lay largely with representatives of the projects. When more projects were added to Follow Through in 1969 and 1970, selection procedures paralleled those used in 1968.

---

[8] Krulee, Organizational Analysis, p. 128.

[9] Krulee, Organizational Analysis, p. 143. D.P. Weikart and B.A. Banet, "Planned Variation From the Perspective of a Model Sponsor," in W. Williams and R.F. Elmore, Social Program Implementation (New York: Academic Press, 1976), pp. 125-158. Elmore, Follow Through: Decisionmaking, pp. 212-216.

So overall, sponsors and projects were "assigned" to one another for a wide variety of reasons. One of the only things that can be said about these pairings was that they were not random and did not result in essentially similar groups assigned to the different sponsor treatments. Some sponsors have more urban than rural sites and others the reverse. The projects associated with one sponsor are mainly in the South, while others have different geographical distributions of projects. Also, three sponsors have only one project each.

Selection of Schools and Children. The selection issue is complicated by the fact that "treatment groups" are by no means entirely defined by projects. The number of schools involved in Follow Through projects ranged from only one in some locations to more than a half dozen in several projects. Also within schools the number of Follow Through classes varied. So not surprisingly, the selection of children to participate in Follow Through varied from place to place too.

Within local education agencies, schools were selected to participate in Follow Through for a variety of reasons. When Follow Through project coordinators were surveyed in 1969-70 as to how Follow Through schools were selected, they gave more then 30 different kinds of answers. Some schools were selected according to characteristics of children and school attendance areas; some were determined by school administrators or other officials; and still others according to individual school characteristics. In two instances schools were chosen at least partly because they had luncheon facilities.

---

[1] SRI, Appendix H, Program Description, Volume IV (Menlo Park, CA: September 1970), pp. 161-63.

Since Follow Through was designed to aid in the continued development of low-income children who previously participated in Head Start, the program naturally established guidelines to ensure that projects served these sorts of children. However, the exact nature of this guide was for a time somewhat unclear, as the 1975 GAO report pointed out:

> The May 1969 memorandum of understanding [between OEO and OE] requires that, with rare exceptions, at least 50 percent of the children participating in each Follow Through grade have preschool experience, preferably a full year, and come from families whose incomes meet Head Start income eligibility criteria at the time of enrollment in Head Start. However, OE's draft program manual, dated February 1969, provided that, with rare exceptions, at least one-half of the children from low-income families in each Follow Through project have a full year of Head Start or comparable preschool experience. According to OE officials, the projects received the manual criteria but not the memorandum criteria.[2]

This ambiguity was cleared up in the 1975 Follow Through Rules and Regulations, and it is not known how much impact the different criteria had on project assignment policies. Nevertheless, it is clear that the proportion of low-income children in Follow Through projects has varied quite widely--from around 50 percent to a full 100 percent.[3]

---

[2] Comptroller General, Follow Through: Lessons Learned, October 1975, pp. 41-2.

[3] According to the 1969 program manual, a family was considered low-income if (1) it was on welfare or (2) its income in relation to family size was below the Office of Economic Opportunity poverty guidelines at the time a child from the family entered preschool or Follow Through. In December 1971 a family of four would have been considered low-income for Follow Through if its annual income was under $3,800. The poverty level has gradually risen over the years. The poverty level as of April 1976 was $5,050 for a non-farm family of four, and in April 1977 was raised to $5,850.

Average income reported in the Abt IV study was $4,998 for cohort III EK analytical sample, $4,459 for the cohort II EF sample, and $4,029 for the cohort II EF sample (Abt IVC, pp. 4, 13, 18). According to the SRI spring 1975 roster data, 86 percent of Follow Through children nationwide met the eligibility requirements for Follow Through comprehensive services. SRI, Feedback Data for Spring 1975 Richmond Virginia (1105) Roster Data (Menlo Park, CA: SRI, undated), p. 6.

The proportion of Follow Through children with Head Start or equivalent preschool experience has been somewhat lower. In spring 1975, around 50% of Follow Through children nationwide had Head Start or equivalent experience.[4] Exact proportions with preschool experience varied widely in different places, however; from 100% in several projects to around 30% in a few instances.[5]

Although generally guided by preschool and poverty criteria, the selection of children into Follow Through has been greatly complicated by a range of factors peculiar to particular projects. The SRI survey of Follow Through project coordinators in 1969-70 revealed nine different types of criteria for selection of children for Follow Through classes;

- Head Start or other equivalent pre-school experience, e.g., local private and public pre-school preparedness programs besides Head Start.
- Low income or economic status (non-differentiated), e.g., all encompassing category for all criteria, e.g., OEO guidelines, welfare roles.
- Selection by larger unit than class, e.g., "all primary children," "entire school," e.g., includes all sites in which program includes all children in the school or primary level.
- Non-low-income, e.g., children selected to maintain possible 10% quota of privileged children as balance to provide realistic school experience, or when they are selected due to an absence of enough deprived children to fill classrooms.
- Volunteers or "advertisement," e.g., in some project areas the political climate or cultural setting is such that children cannot be chosen but must come to Follow Through by means such as advertisement or volunteering.
- Random samplings, e.g., this is most often a subsidiary procedure for picking among equally eligible and deserving Head Start or low-income children. One method is the use of a lottery.

---

[4] According to the spring 1975 roster data 31.6% of Follow Through children nationwide had had Head Start experience, and 17.4% had equivalent preschool experience. SRI, Feedback Data for Spring 1975 Roster Data, p. 5.

[5] For example, only 28% of the FT cohort III EK children in Abt's analytical sample for the Lincoln, Nebraska project had preschool. Abt IVC, p. 114.

- Bilingual, e.g., often accompanies low-income status of Head Start experience, but may be a grounds for selection in and of itself.
- Problem children, e.g., this category includes selection on the basis of criteria that imply a more individual problem such as stimulus deprivation or inability to achieve in a traditional setting.
- Other, e.g., this includes such means as referrals, family situation, etc.[6]

The survey also showed that more than a dozen different kinds of people were responsible for selecting children for Follow Through; ranging from school principals, to Head Start staff to parents to Community Action Agency officials.[7]

In addition to the range of factors influencing the original selection of children for Follow Through, several studies have shown that even more considerations affected selection and assignment in subsequent years. For several years investigators from the Far West Laboratory for Educational Research and Development investigated the assignment of children into Responsive Education Follow Through classes. They found that assignment was influenced by a host of different factors in different localities; including desegregation, special education and parent requests. One Responsive Education program advisor explained,

> If a non-Follow Through classroom gets too large, the teacher will take her ten worst children and send them to Follow Through. If they close a classroom, then Follow Through is a "dumping ground" for the problem children in that class.[8]

---

[6] SRI, Appendix H, Program Description, Vol. IV (Menlo Park, CA: September 1970), pp. 197-99.

[7] SRI, Appendix H, Program Description, Vol. IV, pp. 223-25.

[8] Pierina Ng, "The Assignment of Children to Responsive Follow Through Classrooms," undated draft (Far West Laboratory for Educational Research and Development), p. 5.

Another small-scale study found that in some cases Follow Through has served as an alternative to holding students back to repeat grades. In one project, Follow Through classes were started "in mid-year in the last three years to help Non-Follow Through children who are failing in the regular classroom, and migrant children who move into the area."[9] In another site, the first Follow Through class "was made up of repeaters, slow learners, emotionally disturbed, hard-of-hearing, etc. The classroom looked...like a physical therapy room."[1]

Although such factors have played an important part in the selection and assignment of children into Follow Through in some projects, it is not clear what impact they have had on Follow Through as a whole, nor is there much information available on the extent to which specially selected children have been included in the national evaluation sample. From the service delivery view of Follow Through, the issue is not crucial. Indeed, from this perspective, the assignment of special needs children into the program is a plus. But from the planned-variation-experiment viewpoint, the differential selection of such children into Follow Through has badly complicated matters. In terms of the treatment-control paradigm, it is possible that such assignment policies may have introduced a negative bias into the comparison of Follow Through and Non-Follow Through groups in at least some projects.

---

[9] Jeannette Hargroves, Final Report on the Pupil Assignment and Promotion Study (Cambridge, MA: Huron Institute, June 30, 1974), p. 50.

[1] Hargroves, Final Report, p. 59.

Selection of Evaluation Sample. The treatment-control group model prescribes that both treatment and control groups should be selected in the same way from the same population so that the two groups will be the same at the start of the experiment. The Follow Through reality is again quite different from the ideal.

Selection of Follow Through control groups has been totally dependent upon the selection of Follow Through projects included in the national evaluation. So first, before describing control group selection, let us recap the history of how projects were selected for the evaluation. The evaluation selection issue came up as a major issue at the very first of the SRI planning conference in August 1968.

> The conference participants recognized that the number of projects was not distributed evenly among all sponsors. All agreed that a variety of projects from each sponsor would increase the confidence with which findings might be generalized. A goal of from five to seven projects per sponsor for the eventual evaluation sample was proposed and accepted. (The number of projects in existence in Fall 1968 did not, of course, permit this number to be designated initially.) It was agreed that the group of projects on which to base judgments of sponsor effects should (a) include one or two projects that each sponsor considered exemplary of his approach and (b) exclude (at least initially) projects in which difficulties in implementation were being encountered.[2]

From 1968 to 1972, a variety of factors affected the selection of Follow Through projects into the evaluation sample. Sponsors were encouraged to identify poorly implemented sites for exclusion from the evaluation sample as well as well-implemented sites for inclusion. In the first year of the evaluation, sponsors identified seventeen projects for "certain exclusion" and nine sites for "certain inclusion" on these grounds.[3] Also, the Office of Education picked five projects for certain inclusion and ten for certain exclusion. "From the remaining 65 projects (i.e., ones specified for neither inclusion nor exclusion) SRI selected 35 to satisfy...sampling criteria of frequency and balance across sponsors, regions, and grade streams."[4]

---

[2] SRI, Draft Administrative History, August 1972, p. F-11.

SRI began the 1969-70 evaluation year by pre-testing all entering pupils in 90 projects. After the presentation of the disappointing evaluation results in February 1970, however, the sampling plan for inclusion of projects in the evaluation was shaken up by promulgation of the implementation year policy. Data from 38 of the 90 projects tested in fall 1969-- as well as that from the 49 tested in 1968-69--became ineligible for evaluation because they were drawn from first-year projects. So almost half the projects tested in fall 1970 were excluded from subsequent testing, leaving a total of 46 cohort I projects included in the Emrick report.[5]

The implementation year policy, as well as other considerations, also guided selection of cohort II projects, resulting in 99 eligible cohort II projects tested in fall 1970.[6] Fall 1971 pretesting for cohort III was the broadest testing of any cohort; 81 EK projects and 27 EF projects.[7]

---

[3] Emrick, Interim Evaluation, p. 24.

[4] Emrick, Interim Evaluation, p. 25.

[5] Emrick, Interim Evaluation, p. 27.

[6] Emrick, Interim Evaluation, p. 28.

[7] SRI, "Follow Through Evaluation Data Summaries."

After fall 1971, Follow Through testing was guided mainly by Jane David's sampling plan and particularly the following considerations;

- the pattern of previous testing,
- a policy of including, whenever possible, at least two similar sites per sponsor in cohorts I and II and at least five sites per sponsor in cohort III in order "to maximize the potential for valid comparisons both among Follow Through sponsors and between Follow Through and comparison classes,"
- a policy of making each of the three cohorts as strong as possible in order "to facilitate the determination of how much Follow Through and for whom the program is most effective,"
- special sampling of projects for three sub-studies: a city study, a summer study and a Head Start Planned Variation Longitudinal Study.[8]

Selection of Control Groups for the Follow Through evaluation has been totally conditioned by the selection of projects for the evaluation. Once a project was selected for inclusion in the evaluation, SRI selected

> a set of comparison classrooms in which the characteristics of the children are believed, on the basis of available information, to approximate those of the Follow Through children. In most cases, the comparison classrooms have been located in the same school district as the associated Follow Through project, thus, hopefully controlling for inter-district differences in the basic educational program or in availability of other services.[9]

---

[8] Jane David, Final Report on Sample Selection, June 30, 1973, pp. 1-4.

[9] Egbert, "Follow Through," p. 51. The selection of comparison groups for the RMC cost study was different than described here. See chapter V.

In some cases, however, no satisfactory control groups could be located within the same district as Follow Through schools, so SRI had to find comparisons in adjacent school districts. And in a few instances it was impossible to locate any suitable comparison.

The control group issue was not finished after selection either. Since the control groups were not receiving benefits of Follow Through "treatments," comparison schools had little incentive to spend time and energy having their students tested. As a result, some comparison groups were lost due to "inability... to maintain the cooperation of a control group for data collection periods following the initial one."[1]

The post-hoc selection of control groups and the difficulties in maintaining cooperation of control groups has resulted in a problem familiar to most quasi-experiments of this type. The control groups frequently are not very good matches for the Follow Through treatment groups. Since the Follow Through treatment groups were selected in part on the basis of Head Start experience and poverty, and "control" groups selected afterwards in the same localities, it is not surprising that the control groups are on the average not quite as poor as Follow Through treatment groups nor as largely composed of Head Start graduates. This overall pattern for the Follow Through and control group samples analyzed in the Abt IV report is shown in Table 6. The clear-cut overall trend shown in this figure does not invariably hold when Follow Through evaluation data are examined at disaggregated levels; for example, by sponsor or project. Even so, the mismatch of treatment and control groups is often quite severe at these lower levels.

---

[1] SRI, Draft Evaluation of the National Follow Through Program 1969-71, August 1972, p. II-8.

The mismatch problem at the project level was demonstrated quite strikingly, although indirectly, in a special matching analysis reported in the Abt III study. In the matching study, Abt analysts pooled together all local comparison groups for the cohort II-EK projects. Then for each Follow Through project they selected from this pool a "best match" comparison group. Follow Through projects were paired with best-match comparison groups on the basis of three variables; ethnicity, family income and WRAT pretest score.[2] Of the 52 Follow Through projects included in the study, only twelve were matched by this process with their original local comparison groups.

All of this suggests that Follow Through treatment and control groups were not really drawn from the same population; nor were they drawn in the same manner. Rather it seems that Follow Through treatment and control groups were drawn from overlapping populations, with some of the same sorts of students; but with controls containing more children from relatively rich families and fewer with Head Start or equivalent preschool experience.

* * * * * * * * * * * * *

Selection and assignment of Follow through treatment and comparison groups has been very little guided by experimental design considerations. Although random assignment was tried in the Follow Through summer study, it did not turn out very well in that case, and it has not been tried anywhere else in the national evaluation.

---

[2] The matching procedure was based on a simultaneous matching on the three variables using a least-squares procedure and standardized scores. See Stebbins, Abt III, p. A-55.

Table 6: Overall Comparison of Income and Preschool Experience for Samples Analyzed in Abt IV Report.

	No. of Sites	No. of Children	Mean Income	Percent with Preschool	Fall Entering WRAT Mean Score
Cohort III EK	63				
FT		3705	4792	72	30.2
NFT		2734	5899	45	30.4
Cohort II EK	52				
FT		3367	4660	81	29.7
NFT		2172	5470	57	29.4
Cohort III EF	14				
FT		1346	3836	83	42.5
NFT		850	6101	67	47.3
Cohort II EF	10				
FT		837	3671	71	31.9
NFT		746	4667	54	36.7

Source: Stebbins, Abt IV A, pp. 51, 53, 56, 57.

So overall, the selection and assignment of Follow Through treatments and "control" groups have been guided by a web of interacting administrative, bureaucratic and political considerations. Such factors have operated at several levels, forming a complex hierarchy; with projects selected for Follow Through largely at state and national levels; children to be included selected for participation at the school level; and comparison groups selected by SRI mainly at the level of the school district.

While sponsors' designation of well-implemented and poorly implemented projects, respectively for inclusion and exclusion from the evaluation, may have tended to make evaluation comparisons look good, an apparently more predominant bias has been the post-hoc selection of comparison groups. Since comparison groups generally have tended to be less impoverished than Follow Through treatment groups, overall comparisons would tend to make Follow Through effects look more negative than they really are.

Units and Levels of Analysis

Selection and assignment in the Follow Through planned variation experiment has operated at several levels; at levels of school districts, schools, classes and students. This has contributed to considerable consternation over how this hierarchy of entities should be handled and what should be the unit-of-analysis (UA) for the Follow Through evaluation. Should the evaluation compare individual students' test scores, class average scores, or school or project average scores? The answer is not obvious and indeed has received a variety of different answers during the course of the national evaluation.

The complexity of this problem is illustrated in Figure 7. Within Follow Through there are several cohorts or waves of children and within each wave an entering first and an entering kindergarten stream. Within each cohort-stream there are sponsors; and for each sponsor one or more projects. Within projects there are one or more schools, and within schools there are classes; and in classes are students. How to treat this hierarchy of levels and units is especially complicated in a longitudinal setting like Follow Through since, as suggested in Figure 7, we have to treat a variety of variables and potentially a dozen or more different time points. For the moment, however, let us put aside the temporal dimension and issues of what variables should be examined and consider only how the hierarchy of units and levels has been handled.

For their 1973 Interim Evaluation Emrick and his colleagues used the classroom as their primary unit-of-analysis; that is, they averaged children's data to the classroom level, and then conducted analyses using class level variables. They based their decision on three kinds of considerations, pertaining to: (1) units-of-treatment, (2) variations in levels of measurement, and (3) missing data problems. In the first category, Emrick and his colleagues noted that

> In the Follow Through experiment, pupils are generally grouped into self-contained classrooms...each such classroom represents a homogeneous unit (common teacher, location facilities, and interactions among components)... [and so] classrooms can be viewed as experimentally independent. (p. A-13)

Putting the matter into other words the SRI investigators asserted:

Figure 7: The Hierarchy of Units and Levels in Follow Through

Hierarchy of Levels and Units	Temporal Dimensions and	
	Kindergarten	First Grade
Non Follow Through	. . . . .	. . . . .
Follow Through		
Cohort I		
Cohort II		
Cohort III		
Cohort IV		
Entering First Stream(EF)		
Entering Kindergarten Stream(EK)		
Sponsor 01		
. . . .		
Sponsor 27		
Project 1		
. . . .		
Project i		
School 1		
. . . .		
School j		
Class 1		
. . . .		
Class k		
Child 1		
. . . .		
Child n		

Types of Variables

Second Grade	Third Grade						
	Season						
. . . . .	Fall	Winter	Spring				
			Variables				
			1	2	3	. .	m

> We believe it is reasonable to assume that treatment effects are correlated to a much greater degree within classrooms than across classrooms; that is, what the teacher does with individual pupils or groups of pupils in that classroom but will not necessarily be expected to have any impact on pupils in different classrooms. (p. A-4)

The second argument maintained that the classroom is the appropriate level because,

> Aggregation to classroom-level representation adjusts child and parent variables at the same level as teacher variables. This result is important for purposes of comparability of analysis and interpretation of results, and it enables integration of variables across categories for control purposes. (p. A-6)

The third argument maintained that using the classroom as the UA minimizes problems of missing data. "Even though some or most pupil data within a classroom are lost to attrition, under the assumption that the attrition is random, an unbiased mean can still be estimated for the classroom from the remaining pupil data." (p. A-7)

Using the classroom as the UA also carries certain disadvantages. "On an intuitive level," noted the report, "such a procedure appears to shift evaluation focus from the individual child to the classroom" (p. A-5), while the ultimate goals of the program concern improving the life-chances of the individual child, not the classroom. The second disadvantage is the "reduction of statistical power and precision due to reduced observations." This drawback, the investigators argued, is "somewhat offset by the corresponding reduction in variance presumed to exist at the classroom level" (p. A-5). The third area of concern has to do with the migration of pupils "from year to year across classrooms and teachers" (p. A-7).

Put another way, the difficulty is that the classroom as a naturally occurring unit has a normal duration of only a single school year. The SRI evaluators noted, however, that for the interim evaluation, which focused mainly on comparisons over the span of only one school year, this difficulty was not highly relevant.

The three classroom observation studies solved the unit-of-analysis problem in several different ways. In gathering classroom process data and analyzing patterns in those data, both Stallings and Soar naturally used the classroom as the unit-of-analysis. However, when it came to analyzing the relationships between classroom processes and children's outcome scores, the way to handle the data from these two levels was less obvious. In her 1971-72 report Stallings observed that it would have been desirable to use the classroom as UA but instead employed the child because "to use the classroom as the unit would have necessitated working with extremely small samples" (p. 196). In 1972-73, when her observation study was carried out in 36 instead of only twelve projects, Stallings did use the classroom as the UA. She averaged all scores to the classroom level before conducting partial correlation and multiple regression analyses to investigate the relationship between the classroom level process and outcome data. Soar used a combination of both levels to explore process-outcome relations. Initially using the child as his UA, he calculated regressed gain scores for individual children. Then he averaged regressed gain scores to the classroom level and analyzed the interrelations between outcome and observation scores at the classroom level.

So in the classroom observation studies, the bulk of analyses used the class as UA. Although the child was used as the UA in parts of the first Stallings study and the Soar report, the issue received relatively little discussion in these reports. The focus on the classroom as the unit-of-analysis seemed natural for these classroom observation studies, since most of their data originated from observations on classrooms as units. The issue of units-of-analysis rose to the fore, however, in the 1974 Abt I report.

In this first Abt report, on the 1971-72 data, Cline and his colleagues employed a multiple levels analysis strategy--part of what they called a "cross-validation strategy" (p. V-4). The Abt investigators performed parallel analyses using the child, the class, and the school as units-of-analysis. They attempted to assess specific sponsor effects on a variety of outcome measures alternatively at the pupil, class and school levels of analysis. The Abt investigators advanced the multiple analyses strategy on the grounds that "where results are consistent for parallel questions across the three levels of aggregation, we have enhanced confidence that they represent true effects" (p. V-6). Results showed, however, that consistency rarely prevailed. Effects estimates for individual sponsors were highly inconsistent, rarely agreeing even in direction at all three levels of analysis. And even for the two sponsors whose effects estimates were generally in the same direction across the three levels of analysis, the estimates at the pupil, classroom and school levels varied greatly in magnitude--often by fivefold or more.

There are a number of competing explanations for why the results of the Abt I report were inconsistent across the different levels of analysis. These include different samples at different levels, different sets of variables, and aggregation effects. The issue of how aggregation relations (that is, the relationships between the grouping of individual pre-test and post-test scores to calculate classroom and school averages) was investigated in the context of Follow Through by Walt Haney of the Huron Institute.[3] Using a single data set he showed that in strictly parallel analyses the natural aggregation or grouping relations in and of themselves can cause markedly different estimates of sponsor effects from analyses using the individual, class and school as units-of-analysis.

The problem of how to handle different units and levels of analysis recently has been examined in a broader context by a variety of investigators.[4] The most startling conclusion from investigation of these issues came from Lee Cronbach. In a study entitled Educational Effects of the Individual and Group Levels; the Crucial Importance of Hierarchical Interpretation, Cronbach wrote:

---

[3] Units of Analysis Issues in the Evaluation of Project Follow Through or There Must Be Heresy in This Someplace (Cambridge, MA: Huron Institute, September 1974). The paper details the range of arguments advanced to justify using different analysis units.

[4] At the 1976 American Educational Research Association convention, a special symposium was held to consider issues of data aggregation and units-of-analysis. At the conference, papers were presented by Leigh Burstein of UCLA, Michael Hannan of Stanford University, and Linda Glendening and Carlyle Maw of the National Institute of Education, among others.

> I find the two chief implications so startling that I hesitate to put them forward.
>
> 1. The majority of studies of educational effects--whether classroom experiments, or evaluations of programs, or surveys--have collected and analyzed data in ways that conceal more than they reveal. The established methods have generated positively false conclusions in many studies, and have led to claims of statistical significance for numerous chance results.
>
> 2. The traditional research strategy--pitting substantive hypotheses against a null hypothesis and requiring statistical significance of effects--can rarely be used in educational research. Samples large enough to detect strong but probabilistic effects are likely to be prohibitively costly.[5]

Cronbach's implications are startling indeed and will surely draw further attention to the issue of units and levels of analysis.

Nevertheless, for present purposes two caveats concerning the inference of program effects in quasi-experiments like Follow Through seem warranted. The first is that a relationship estimated at one level of analysis can be inferred to another level (for instance, to say that a correlation or regression relation among classroom-level variables holds at the individual student level) only with great caution. Indeed, this kind of inference was once labeled the "ecological fallacy," and later and somewhat more generally termed the fallacy of the wrong level.[6]

---

[5] Lee J. Cronbach, Educational Effects of the Individual and Group Levels; the Crucial Importance of Hierarchical Interpretation. First Report to the Spencer Foundation. School of Education, Stanford University, December 1975, Foreward. In a 1976 revision of this report, Cronbach repeats these two conclusions with only minor changes of wording. Research on Classrooms and Schools: Formulation of Questions, Design, and Analysis. Occasional Papers of the Evaluation Consortium, Stanford University, Stanford, CA, July 1976.

[6] "Ecological fallacy" was the term used by W.S. Robinson in 1950, and the more general "fallacy of the wrong level" was identified by John Galtung. See Haney, Units of Analysis, pp. 14-19.

However, as Michael Hannan has pointed, out, the notion of a fallacy may be useful initially to point out a problem, but it also may overstate the case.[7] Relationships prevailing at one level of analysis may suggest the nature of relationships at other levels, but the correspondence in relationships at different levels should not be assumed without investigation of the aggregation relations between levels.

Second, if we are interested in the effects of the program on individual students, investigations ought to use the individual as the unit-of-analysis, or at least to conduct anlayses at other levels only in conjunction with individual-level analyses. A variety of arguments have been advanced in favor of using the class or the school as the unit-of-analysis. Yet these are not terribly persuasive because they allow technical issues to hold complete sway over more substantive issues regarding program impact on individuals. If the goal of the Follow Through program is to have an impact on individual children, then this should guide the choice of a unit-of-analysis.[8]

The trend in the Follow Through evaluation has clearly been away from higher units or aggregations of analysis and toward a focus on the individual child. In the Abt II, III and IV reports the child was the primary UA.

---

[7] M. Hannan, Aggregation and Disaggregation in Sociology (Reading, MA: Lexington, 1971), pp. 9-10.

[8] For more on this argument see Haney, Units of Analysis. The only exception to this argument might be the sort of situation suggested by the GAO, wherein projects (or other entities above individual students) are randomly assigned to treatments. However, this is rarely the case. Also, it should be noted that this conclusion is not, by any means, universally accepted. In fact, Cronbach seems to advocate a hierarchical analysis beginning with analyses at higher levels, like the school, and working downward to class and individual level analyses.

This downward trend will be examined more closely in the section of this chapter on ANCOVA. Before moving to this detail, however, we outline some of the other issues--like the idea of a treatment and the problem of attrition--which have been woven into discussions of the appropriate unit-of-analysis for the Follow Through evaluation.

The Follow Through Treatment: Specification and Verification

> The idea of planned variation may have been totally logical at the national level, but at the sponsor level it was mystifying. The meeting in Washington of prospective sponsors to present their various orientations to curriculum, the idea that evaluation could be done by a third agency,...and the assumption that each sponsor had a complete package to present--all created a feeling of bewilderment and even of madness.
>
> \- David Weikart and Bernard Banet[9]

The treatment-control experimental paradigm suggests that the treatment to be tested is well-defined, so that it can easily be determined whether or not the treatment has been administered. The evaluator needs to be able to verify whether or not the treatment has been implemented; otherwise the experiment may be testing only a pseudo-treatment instead of the real thing. On both counts--specification and verification of implementation--the evaluation of the Follow Through curriculum models falls considerably short of the elegant simplicity of the paradigm. This section treats the implementation issue in three parts: how implementation has been handled in the impact or outcome studies, how it has been treated in the classroom observation studies; and how concern has arisen over the diffusion of Follow Through "treatments" to the control groups.

---

[9] David P. Weikart and Bernard A. Banet, "Model Design Problems in Follow Through," in A. Rivlin and P.M. Timpane (eds.), Planned Variation in Education (Washington, D.C.: Brookings, 1976), pp. 61-77.

Implementation in the Impact Studies. As Weikart and Banet suggested, Follow Through sponsor models were far from complete packages or well-specified treatments at the start of the Follow Through program. Part of the problem was the lightning-fast beginning to Follow Through. The period of conception for the planned variation experiment was a mere nine months--from an idea in Washington in December 1967 to implementation in schools throughout the country in the fall of 1968. But the issue of model specification was not simply a start-up problem. Curriculum models have continued to change and evolve throughout the course of Follow Through. Six sponsor models have changed in institutional affiliation during the course of the program and in at least "seven others the involvement of the creators had been either greatly reduced or lost entirely."[1]

These sorts of changes undoubtedly hindered the push for specification of sponsors' models, but also there has been a more fundamental problem. This is that the very idea of a mechanistic, standardized "treatment" is foreign to the approaches of some Follow Through sponsors. They hold that it is inappropriate, if not impossible, to prescribe what a school or community needs without considering local conditions and needs. EDC has been one of the sponsors sharing this view:

---

[1] Garry L. McDaniels, "Evaluation Problems in Follow Through," in Rivlin and Timpane, p. 50.

> ...we, at EDC, believe that the educational approach should always be evolving and that as part of this Open Education philosophy and approach to learning, our goals and objectives are always developing and changing in order to meet the specific and immediate needs of our Follow Through communities.[2]

Even if models are well-specified, they cannot be evaluated unless they are actually implemented. Indeed, if the evaluator cannot be sure that a treatment has been implemented then the evaluation may be merely a pseudo-experiment instead of a real test of alternatives. Concerns like this have given rise to much talk and writing about the implementation of sponsors' curriculum models. For example, the 1972 review panel suggested that the implementation issue should be investigated in three parts:

1. The extent to which the nature of actual treatment is known or knowable.
2. The extent to which treatments (vs. sponsor intentions) can be transported or implemented at sites other than those in Follow Through.
3. The extent to which models are maintained subsequent to withdrawal of Federal support.[3]

The review panel recommended "further research and evaluation" into each of these areas, as well as into "an explanation of how implementation works; specifically how does an essentially R&D model become translated into practice in a 'real world' setting."[4]

As has happened so often in Follow Through, the questions far outstripped the ability of evaluators to produce firm answers. Emrick and his colleagues at SRI recognized the implementation issue, but could do little about it. They disclaimed that, "we now have only qualitative distributions (of treatments) corresponding to sponsor descriptions of their respective models, and no assurance that variability in treatment is any less across projects within a given sponsor than across projects between sponsors" (p. A-12).

---

[2] Grace R. Hilliard, "Overall Goals and Objectives of the EDC Follow Through Program," letter to Dr. Mary Kennedy, February 6, 1976.

[3] OE, Second Report, pp. 26-27.

[4] OE, Second Report, p. 28.

Abt investigators also took up the implementation issue in the first phase of their evaluation work. They asked, "What type of inputs can reasonably be expected to contribute to the construction of scales of sponsor implementation?" (Abt IA, p. III-4). In pursuit of an answer, the Abt evaluators visited nine Follow Through projects in New York and Philadelphia. Reporting in Volume IB of the Abt series on Follow Through, the investigators documented a plethora of local conditions which complicated the implementation of curriculum models in the two cities: a "bitter teacher strike," bias among administrators toward a particular model, overlapping of staff for different models causing mixing of the models, and staff turnover, among others. The Abt effort to investigate the implementation issue fell completely short of the initial query concerning the construction of scales of implementation. Instead Cline and his colleagues concluded only,

> ...the "real life stories" do offer some clues for deeper insights into the human variables in program delivery than can be obtained from the most sophisticated quantitative analyses (p. III-35).

In their second report the Abt investigators did not directly address the implementation question; but in conclusion they again stressed its importance. They wrote: "For most of the sponsors, site dynamics appear to mediate or even to attenuate their effects" (p. VII-1). The Abt evaluators lamented the fact that site variations make each sponsor's effects "difficult to interpret" and also impose "statistical restrictions on the interpretations" (p. VII-9). Cline and his colleagues recommended that the problem of variability "can only be unraveled by detailed studies of the site level" (p. VII-10). As a "major recommendation of this second annual evaluation of Follow Through," they advocated initiation of a "directory of effects associated with types of sponsors under types of site conditions" (p. VII-10).

As recounted in chapter 1, one of the major issues set out for the second phase of the Abt Follow Through evaluation was again how implementation of Follow Through models could be assessed. The classroom observation data gathered in the Stallings and Soar studies was not suitable for the Abt analyses because they dealt with a much smaller number of projects than Abt had to handle. So the Abt analysts investigated two alternatives for assessing implementation. One was to have sponsor staff members, Abt staff, and OE project officers rate the degree of implementation in different projects. In early 1975, Abt surveyed these people "who know about the realities of the Follow Through sites." Each was asked, "From what you know about this site, how much of the sponsor's treatment would be estimated as reaching its intended recipients."[5] People were asked to rate implementation on a 0 to 5 scale, ranging from "practically no implementation" to "practically perfect implementation."

---

[5] Richard B. Anderson, memorandum "Implementation and Comparability" to "People Who Know About the Realities of Follow Through Sites," 19 March 1975.

The rating approach to implementation soon short-circuited. Abt was unable to get ratings from some sponsors and project officers. Also, the reliability of implementation ratings turned out to be highly questionable. While the Abt staff and OE project officers seemed to have "a fairly common perspective," the match of their ratings with those of sponsors was not so good. In one case even the rank ordering of ratings was "really out of whack."[6] This naturally raised questions about the validity of the implementation ratings.

> Perhaps we are rating the politics of the community and the sponsor is rating classroom processes or some other dimension. While [one sponsor] refused to help us rate their sites on any dimension, they did agree that no single dimension would define implementation. Those which came up were classroom practices, parent-community involvement, political problems. They also felt that classroom practices had to be analyzed at the teacher level.[7]

As a result of these problems, the rating approach to assessing implementation was scrapped and the Abt analysts turned to data from the parent interview and teacher questionnaire data as a source of implementation information.[8] The parent and teacher data had been analyzed at Abt with both factor and cluster analysis, and in the draft of the Abt III report, the variables derived from the parent and teacher data were described as empirical estimates of children's home and school learning environments. When sponsors reviewed the draft Abt III report, several objected to this approach. For example, one maintained that "parents' perceptions of the program are so personalized and impressionistic that they cannot be termed empirical estimates."[9] At least partially as a result of sponsors' objections, the parent and teacher variables were linguistically devalued in the final Abt III report. The variables from the parent and teacher data were termed merely "indicators of acceptance of model tenets" (Abt III, p. A-24).

---

[6] Nancy Ames, memorandum "Implementation Study" to Richard B. Anderson, undated [1975].

[7] Ames, memorandum "Implementation Study."

[8] The broader history of how parent and teacher data have been gathered and analyzed through the history of the Follow Through evaluation is recounted in chapter IV.

In their final report, the Abt evaluators backed away even further from the implementation issue. In a section entitled "lack of specificity of treatment" they warned,

> There has been no adequate assessment of whether the sponsors' evolving programs were implemented. We have, therefore, had no way to address the research issue, "Do differences of implementation of an educational strategy have any relationship to differences in impact?" (Parent and teacher data related to implementation are used as descriptive material only.) (Abt IVA, p. 10)

Classroom Observation Approaches to Assessing Implementation. All five of the major Follow Through impact studies--the Emrick, and the Abt I, II, III and IV reports--discussed the implementation issue and its importance to interpreting effects estimates. But none of these Follow Through evaluators--neither Emrick et al, nor Cline and company nor Anderson, Stebbins and associates--could make much progress on the problem. Their reactions to this frustration were similar--they wanted more information. Emrick proclaimed the "need for more precise treatment definitions and descriptions" (p. 315); Cline and company called for "a directory of effects associated with types of sponsors under different types of conditions" (Abt IIA, p. VII-10), and in their summary of the Abt IV report, Stebbins and associates advised,

---

[9] Elizabeth Gilkeson, letter to Director Follow Through Evaluation, Abt Associates, February 18, 1976, enclosure.

> Future evaluations that aspire to explain the outcomes of compensatory intervention will do well to investigate thoroughly the local circumstances, attitudes and activities that clearly decide whether a compensatory education program can work or not. (Abt IVA, p. xi)

The more-information-needed remedy sounds plausible--but the experiences of Soar and Stallings, who both tried to illuminate the implementation issue, raise doubts. Soar experienced the issue only indirectly. He investigated how Follow Through programs differ in terms of classroom observations instead of asking whether classroom observations corresponded to particular models. He used six classroom observation instruments to gather information on 289 Follow Through and Non-Follow Through classrooms. Together the instruments yielded about 500 bits of information on each class, even though they were used only one day per class. Altogether, this yielded tens of thousands of pieces of information. In order to sort through this massive amount of data, Soar used factor analysis to identify common dimensions of variability. He found that certain of his factor variables did distinguish between classrooms associated with different sponsors. Yet when he looked at more than one such variable at a time, that is, when he considered patterns of factor scores, he found that the "results do not appear to suggest a very high degree of consistency of classroom behavior within program" (p. 158). Soar noted, however, that his was a "relatively fine-grained analysis" and his "measures were not chosen to represent the objectives of the sponsors, but rather to identify differences between teachers" (p. 165). Therefore, it might have been argued that the diverse patterns revealed by Soar's measures were due to aspects of teacher behavior irrelevant to sponsors' models and hence to the treatment implementation issue.

In her 1971-72 study, Jane Stallings studied the implementation issue in four phases. First, she described each Follow Through model from a theoretical point of view. Second, based upon the SRI classroom observations she described what a day in each model's classroom was like.[1] Third and in order to analyze systematically how the classroom observations conformed to sponsor specifications, she constructed implementation scores for each sponsor. Fourth, she factor analyzed 65 observation variables and from these factor scores constructed profiles on each of the nine resulting factors for each of the twelve sponsors in her 1971-72 study.

The implementation issue was tackled even more directly in the 1972-73 Stallings study. In this study, Stallings' number one question was:

> Are the observed classroom programs consistent with their sponsors' stated intentions? That is, does the model show a relatively high frequency of occurrence of those elements of the program that the sponsor rated as important? (p. 6)

Stallings' initial problem was specification. There were no comprehensive descriptions of sponsors' models in terms of classroom behaviors. Without such descriptions, it would be impossible to assess how well models had been implemented. So in the spring of 1973 Stallings visited "ideal first and third grade classrooms specified by each sponsor," in the hope of defining models in behavioral terms by observing ideal model classrooms.

> What would be desirable for a closely controlled experiment would be an explicit statement by sponsors of what proportion of the time critical variables would occur in an ideally implemented classroom. However, elements in a classroom are not like those in a test tube. We have not yet learned to predict the amount of individualized attention, or feedback

---

[1] It is worth noting again that among all the mounds of materials resulting from the Follow Through evaluation these descriptions by Stallings are unique. They comprise the only direct empirical descriptions of what life is like in the classrooms of different sponsors. See Stallings 1971-72, pp. 79-193.

> that a group of children need to meet specific goals. Each classroom group is made up of individuals, and individuals are most likely to need different rates of feedback or individualized instruction for maximum growth. Thus, even the most specific models such as University of Oregon and University of Kansas will alter the rate of reinforcement depending upon the need of the child ..We simply could not come up with a recommendation that would specify the rate of questions or reinforcement to be expected from even the most structured models. (p. 95)

In consequence, the attempt to specify ideal model types directly did not work. So Stallings used a roundabout approach to the assessment of implementation. From the classroom observation instrument, Stallings and other SRI investigators selected sets of variables "intended to describe representative classroom elements for each sponsor's model." After surveying sponsors as to which of the observation variables would be considered critical to the sponsor's model, they developed a list of critical variables with which to assess implementation. These variables are listed in Table 7. Unfortunately, Stallings does not describe the exact sequence of steps leading to the list of critical variables. Still, the list suggests a lot about what different sponsors rate as important. Also, it shows much about the implementation issue. Of the 55 variables, three were judged critical to all seven models included in the study; and twelve were judged critical to at least five different models. Another twelve of the variables were judged uniquely critical to only one model. The list points out that sponsors' models are not unitary; some aspects considered critical for first grade are not so rated for third grade, and for other variables the reverse. In addition, the totals for the seven sponsors raise doubts about the comprehensiveness of the list. Can seventeen variables--even 29--

No.	Variables*** Description	Far West Labs	University of Arizona	Bank Street	University of Oregon	University of Kansas	High Scope	EDC	No. of times judged critical
24	Child selection of seating and work groups	X	X	X			X	X	5
25	Games, toys, play equipment present	X	X	X			X	X	5
39	General equipment, materials present	X				X		X	3
65	Guessing games, table games, puzzles		X			X	X		3
66	Numbers, math, arithmetic	X	X	X	X	X	X	X*	7
67	Reading, alphabet, language development	X	X	X	X	X	X	X*	7
70	Sewing, cooking, pounding		X				X		2
71	Blocks, trucks						X		1
74	Practical skills acquisition	X**		X					2
83	Wide variety of activities, over one day	X	X	X			X	X	5
86	Teacher with one child	X	X	X			X		4
87	Teacher with two children			X					1
88	Teacher with small group		X	X	X	X	X		5
92	Aide with one child	X		X					2
94	Aide with small group		X		X	X	X		4
114	One child independent	X	X	X				X	4
115	Two children independent							X	1
116	Small group of children independent	X		X				X	3
239	Math or science equipment/Academic Activities	X		X			X	X	4
240	Texts, workbooks/Academic Activities				X	X			2
343	Child to adult, all verbal except response			X					1
344	Individual child verbal interactions with adult	X	X	X	X	X	X	X	7
350	Child questions to adults	X	X	X			X	X	5
363	Child group response to adult academic commands/requests or direct questions				X				1
372	Child presenting information to a group			X			X		2
375	Adult instructs an individual child	X						X	2
376	Adult instructs a group				X				1
390	Adult task-related comments to children						X	X	2
394	All adult acknowledgment to children	X		X	X	X	X		5
398	All adult praise to children		X		X	X			3
412	Adult feedback to child response to adult commands/requests, questions				X	X			2
420	Adults attentive to a small group	X			X		X		3
421	Adults attentive to individual children	X	X	X		X	X		5
423	Positive behavior, adults to children	X	X	X					3
435	Total academic verbal interactions				X				1
438	Adult communication or attention focus, one child	X		X		X	X	X	5
440	Adult communication or attention focus, small group				X		X		2
444	Adult movement	X							1
450	All child open-ended questions							X	1
451	Adult academic commands/requests and direct questions to children				X	X			2
452	Adult open-ended questions to children	X	X	X			X	X	5
453	Adult response to child's question with a question						X	X	2
454	Child's extended response to questions		X	X					2
456	All child task-related comments	X	X				X		3
457	All adult positive corrective feedback	X			X	X			3
460	All child positive affect	X	X					X	3
469	All adult reinforcement with tokens					X			1
509	Child self-instruction, academic				X*				1
510	Child self-instruction, objects			X			X	X	3
513	Child task persistence			X		X	X		3
514	Two children working together, using concrete objects						X		1
515	Small group working together, using concrete objects			X			X		2
516	Social interaction among children	X		X				X	3
574	Child movement	X						X	2
599	Child self-instruction, nonacademic	X	X				X		3
	Total number of Critical Variables	28** 27*	21	27	16** 17*	17	29	20** 22*	

---

*Third grade only.

**First grade only.

***Variable numbers refer to items on Stallings' classroom observation instruments.

Source: adapted from Stallings 1972-73, p. 99.

adequately capture all the critical aspects of a sponsor's model? What's more, the manner in which the list was created raises a question which harkens back to the parent protest at Denver: Are evaluators measuring that which is important, or treating as important that which can be measured?

Even if one accepts the list as comprehensive, and as measuring that which is important, it still leaves problems about how to use it. How can the list be used to assess the degree of model implementation? Stallings answered the question as follows. She simply standardized classroom scores on the variables in terms of how Non-Follow Through classes scored on them, and summed across all the critical variables for each Follow Through class to derive an overall implementation score for each class. The aggregation process summed together scores on observation items as diverse as variables #39, general equipment; #66, numbers, math, arithmetic; and #469, adult reinforcement with tokens. Total implementation scores for each site were then calculated by averaging classroom total implementation scores for each site. In turn, site-level implementation scores were used to calculate total implementation scores for each sponsor. Using t-tests, Stallings then compared sponsor implementation scores with corresponding scores for Non-Follow Through classes. She concluded,

> No simple statement, such as that all seven Follow Through sponsors have succeeded in implementing their model in the sites which were studied, can be made. Rather, conditional statements must be made. While all sponsor classroom mean scores for the first and third grades are significantly different from the Non-Follow Through mean score, all sponsors except Far West Lab and University of Kansas and High/Scope have some classrooms which have scores below

> the Non-Follow Through mean. In both grades the University of Kansas implementation scores were higher than those of other sponsors. (p. 195)

The findings are intriguing. Yet they raise doubts about the roundabout approach to assessing implementation and the real meaning of Stallings' implementation scores. For example, it seems that the high implementation scores for Kansas may have resulted, at least partially, from the fact that the set of critical variables identified for Kansas was far smaller than those for most other sponsors and that it contained the most unusual of any of the critical variables--variable #469, adult reinforcement with tokens. Stallings' implementation study can be questioned in a number of other respects--for instance with regard to observer bias and reliability, and the unusual standardization procedures employed.[2] Nevertheless, Stallings' work represents the most comprehensive evidence so far available that classrooms associated with Follow Through sponsors generally do differ from Non-Follow Through classrooms in at least some of the ways intended by sponsors. At the same time, however, Stallings' work raises an intriguing question about how these differences came about.

In an effort to identify "elements in the sponsors' in-service teacher training programs which were effective in the implementation process," Stallings compared classroom implementation scores to various items from a teacher questionnaire relating to teacher "training, teacher experience, education and satisfaction with the sponsor's model."

---

[2] For a brief description of Stallings' standardization procedures, see the bibliography. On the possibility of observer bias, or "the confounding of observers with models" see commentary by Laren B. Resnick and Gaea Leinhardt in Monographs of the Society for Research in Child Development, Vol. 40, No. 163, Nos. 7-8, December 1975. This volume of the Monographs was devoted entirely to Stallings' Follow Through classroom observation study. It should be added that the brief description given here focuses merely on one of the three such parts of Stallings' study.

> Unfortunately the items on the questionnaires were inadequate to yield the information necessary to understand the sponsors' teacher training strategies. Partial correlations were computed between implementation scores and items from the teacher questionnaire. No significant correlations were found. (p. 191)

The problems of treatment specification and verification of implementation are not necessarily solved by more information. Both the Soar and Stallings reports suggest that more information may be gathered; but it may not mean more clear cut answers. The Soar and Stallings studies are remarkable efforts; both recount creative solutions to extremely difficult issues. But both skirt around the treatment issues and raise more questions than they answer. What are the Follow Through model treatments? How are they implemented? Isn't there a way to verify whether a particular model has been implemented in a particular classroom without comparing it to other classes defined only as non-treatment? Are there no means, other than highly abstract factor scores or standardized implementation scores, for describing model implementation?

As David Weikart and Bernard Banet observed, "it is not necessarily the methodology that is at fault even in 'quasi-experiments' like Follow Through...What is needed may be a clearer conceptualization of models, outcomes, and intervening processes, as well as better educational treatments for the researcher to evaluate" (pp. 70-71).

Control Groups and Diffusion of Treatment. A problem basic to all of the efforts to assess implementation as well as to those aimed at estimating effects is the nature of the "control" group, and the "treatment" it receives. The ideal embodied in the treatment-control paradigm is that control groups receive no treatment. Yet the "no treatment" condition clearly has not held for the FT control groups. All of the control group children have been receiving some sort of schooling. Since the control groups were chosen on the basis of similarity with Follow Through projects, including the low income criterion, most NFT control schools were in fact receiving special services under Title I of ESEA.[3] This issue has not been investigated comprehensively, but it was considered in the Abt II report. The Abt analysts investigated the nature of services provided in the FT and NFT schools included in their analysis. Of the roughly 400 schools surveyed they found that 92% of FT schools and 85% of NFT schools were receiving Title I funds. Also they discovered that NFT schools were providing many of the same kinds of specialized services as delivered in FT.

> This finding...poses some serious problems for the interpretation of FT/NFT contrasts. FT is not being compared to a "regular" school program, but rather to supplementary programs over and above the basic level of effort provided by the local educational agency. The quality of such programs varies widely, suggesting that additional information on the comparison programs would greatly enhance our understanding of pupil results (Abt IIB, p. II-60).

---

[3] A 1976 survey by the National Institute of Education revealed that an average of 90% of the schools classified by local education agencies as eligible to receive Title I funds did in fact receive them. National Institute of Education, Evaluating Compensatory Education. An Interim Report on the NIE Compensatory Education Study (Washington, D.C.: NIE, December 30, 1976), p. xvi.

So the Follow Through models are not in reality being compared to a no-treatment control nor even to regular schooling, but instead to alternative compensatory education programs.

A related problem is that some aspects of the Follow Through treatment may have "diffused" or seeped over to the NFT comparison groups. So in at least some cases the evaluators may unwittingly actually be contrasting FT treatments not even with an alternative compensatory program, but instead with a partial FT treatment. The nature of this problem (and the different light in which it appears from non-experimental perspectives on FT) was shown very clearly in an OE status report on Follow Through in 1974. Entitled "Follow Through Evaluation Report," the document was submitted to a subcommittee of the U.S. Senate Committee on Appropriations. It noted,

> Several communities have made the initiative to further extend Follow Through programs in their school systems.
>
> For example, in Tacoma, Washington, the Follow Through model which has operated in seven Follow Through sites has been incorporated in all the city's Title I kindergarten classrooms.
>
> Another example is that several Indian tribes intend to extend Follow Through activities with funding under the Indian Education Act to non-Follow Through funded classrooms at the early childhood education level. These include the Rosebud reservation in South Dakota, the Hope [sic] tribes in Oraibi, Arizona, and the Eastern Band of Cherokee in Cherokee, North Carolina.[4]

The diffusion problem has concerned Follow Through evaluators for several years, but most have been able to do little more than point to it as a problem.[5] From the classroom observation studies, it is apparent that there is little way to assess the diffusion issue definitively. Virtually all the activities occurring in Follow Through classrooms also occurred in Non-Follow Through classrooms. Moreover, the kind of schooling for Non-Follow Through groups varies widely. This was shown quite dramatically in an analysis conducted by Jane Stallings and reported in her 1972-73 report. In order to test how well classroom observation data discriminated among classrooms of different sponsors, Stallings empirically classified classrooms with sponsors. The analysis aimed to test "how reliably a classroom can be identified with the correct sponsor or, conversely, how often it is confused with the program of another sponsor" (p. 200). The analysis was conducted in four parts, using two sets of observation variables separately for first and third grade classrooms. Via these analyses Stallings found that classrooms could be correctly identified with sponsors in around 80 percent of the classifications (although misclassifications were less than 10 percent for both Kansas and Oregon classrooms).[6] Stallings also applied her classification functions "to classify the non-Follow Through classrooms in order to examine the relative degree to which the processes in non-Follow Through classrooms are similar to the various

---

[4] U.S. Congress, Senate, Hearings before the Committee on Appropriations, May 29, 1974, p. 2362.

---

[5] For example, McDaniels in "Evaluation Problems in Follow Through" (p. 51), termed the problem "seepage," and the Abt analysts in their final report on Follow Through wrote, "In some instances, educational aspects of the Follow Through treatment may have diffused from Follow Through into non-Follow Through classrooms, thus contaminating the comparison groups with unknown portions of the treatment (Abt IVA, p. 10).

[6] Stallings 1972-73, pp. 221-31.

Follow Through programs" (p. 227). Of the 140 Non-Follow Through classifications made, at least ten Non-Follow Through groups were assigned to each of five different Follow Through sponsors.[7] In other words, as evidenced in Stallings' classroom observation variables, the Non-Follow Through "treatments" appear to vary nearly as widely as the different Follow Through models themselves.

Other evidence on the control treatment and diffusion issues is found in the teacher questionnaires. In the spring of 1972 NFT teachers were asked "From what you know about Follow Through programs, are there any aspects of these programs that are similar or identical to what you do in your classroom?" Although responses to this question were meager (only 72 NFT teachers answered it), more than 10% of the responses were yes.[8] In later years, teacher questionnaire data revealed that there has been some diffusion of personnel from FT to NFT groups. In each of the 1973, 1974, and 1975 teacher surveys, around five percent of NFT teachers indicated that they had previously taught in a Follow Through approach.[9]

In sum, it is clear that NFT "control" groups approximate the ideal of the randomized treatment-control group experiment far more in the eyes of evaluators than they have in actuality. It is clear that the NFT groups did not experience a "no-treatment" condition but in fact received widely varied sorts of schooling, in some cases similar to the sort of instruction received by Follow Through children. It is almost certain too that most of the schools in the Non-Follow Through comparisons received assistance under Title I of the Elementary and Secondary Education Act of 1965. Beyond this, however, relatively little is known about the "treatments" received by the so-called control groups. What is clear is that the FT treatments are up against a much tougher standard of comparison than a real no-treatment control group. Indeed, there is some irony in the fact that several of the Follow Through sponsor models are less directly aimed at the sorts of outcomes encompassed in the Follow Through test battery than are the Title I programs with which they apparently have been compared.

---

[7] Stallings *1972-73*, p. 224. The only sponsor to which less than ten NFT groups were classified was Kansas. This anomaly apparently resulted from one particular variable associated with the Kansas approach--adult reinforcement with token--which occurs hardly anywhere outside of Kansas classrooms.

[8] Stallings *1971-72*, p. 54. From her consideration of these data Stallings observed that "the degree of diffusion between Follow Through and Non-Follow Through varies across sites."

[9] SRI, *Feedback Data for Spring 1975*; Teacher Survey, p. 3, and corresponding volumes for 1974, and 1973.

> In discussions associated with the preparation of the Education Amendments of 1974 (P.L. 93-380) the Report of the House Committee on Education and Labor stated "that local educational agencies should give priority attention in operating Title I programs to the basic cognitive skills in reading and mathematics and to related support activities to eliminate physical, emotional or social problems that impede the ability to acquire such skills."[1]

---

[1] Office of Education, Office of Planning, Budgeting and Evaluation, *Annual Evaluation Report on Programs Administered by the U.S. Office of Education FY 1975* (Washington, D.C.: Office of Education, undated), p. 82. See pp. 20-31 of House Report No. 93-805. "Both House and Senate discussions (see Senate Report No. 93-763, pp. 30-31) recognized that such an assertion was not intended to preempt the prerogatives of local authorities to give priority to other areas if such emphases were required to better meet the needs of disadvantaged children."

Duration of the Treatment and the Evaluation

> [T]he Follow Through treatment provides a unique opportunity for studying children over a period of three or four years. But the question inevitably arises, how many years are enough? Even in Follow Through, with its longitudinal design and successive replications, four years is attacked as too restricted a treatment period; stopping the treatment at the end of the third grade as a time "when the problems are just beginning"; or any kind of testing that is not continued into the upper elementary grades and junior high as unrepresentative of a true estimate of the impact.
>
> -Garry McDaniels[2]

Another aspect of the treatment-control paradigm which has spawned dispute during the course of the Follow Through evaluation is the temporal one. Interim Follow Through evaluation analyses have estimated Follow Through treatment effects at the end of kindergarten, first grade or second grade. But the prevailing assumption has clearly been that since the Follow Through program lasts through the third grade of elementary school, the definitive effects estimates should be based on spring third grade test scores. This thinking is based on the assumption that Follow Through is a four-year treatment for entering-kindergarten stream children and a three-year treatment for children in entering-first streams. In this model, the temporal boundaries of the treatment are fall kindergarten (or fall first grade for the EF streams) and spring third grade. These points are usually assumed to be the appropriate pre-test and post-test time, respectively, for the Follow Through evaluation. Both boundaries have, however, been challenged, and so we need to consider the arguments advanced in favor of alternative bounds for the Follow Through treatment and evaluation. First discussed is the start of the treatment and the appropriate pre-test time, and then the end of the treatment and the appropriate post-test time.

---

[2] G. McDaniels, "Evaluation Problems in Follow Through" in A. Rivlin and P.M. Timpane, Planned Variation in Education (Washington, D.C.: Brookings, 1975), p. 49.

Start of the Follow Through Treatment and Time of Pre-test. The official description of Follow Through in the *Federal Register* states that Follow Through is designed to assist the overall development of children in kindergarten through third grade. However, as the official description also notes, Follow Through is designed to amplify the gains made by children in Head Start and similar preschool programs. So Follow Through can be viewed as only part of a combined Head Start-Follow Through treatment. From this perspective the appropriate pre-test time is not entry into kindergarten, but entry into preschool.

This view--of a combined Head Start-Follow Through treatment--was part of the motivation for the Office of Child Development which administers Head Start to begin a special Head Start Planned Variation (HSPV) study in the fall of 1969. The HSPV program was a three-wave longitudinal study designed to assess the relative impacts of a variety of preschool curricula. The program was carried out in Follow Through sites, and employed preschool curriculum models devised by eight of the Follow Through sponsors. Data from each of the three HSPV cohorts have been separately analyzed,[3] but what concerns us here is the notion of a combined HS-FT treatment.

---

[3] L. L. Dittman, et al., *Study of Selected Children in Head Start Planned Variation, 1969-70* (Menlo Park, CA: Stanford Research Institute, March 1971).
M. S. Smith, *Some Short Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation -- 1970-1971* (Cambridge, MA: Huron Institute, January 1973).
H. I. Weisberg, *Short Term Cognitive Effects of Head Start Programs: A Report on the Third Year of Planned Variation. 1971-1972* (Cambridge, MA: Huron Institute, June 1974).

The original idea behind HSPV was that after Head Start, HSPV children would enter Follow Through and be tested while in Follow Through. Children's test records could then be linked and would, went the idea, provide the basis for evaluating the combined preschool and early elementary experiences of the Planned Variation children. Things did not work out as planned for a variety of reasons. At the federal level, Follow Through and Head Start are administered by different agencies (OE and OCD) and coordination between them has not always been smooth. At the local level, Head Start and Follow Through often were not coordinated so that HSPV children went into and were tested in Follow Through. In addition, there has never been any traced comparison group of children who did not attend Head Start or Follow Through, but were still tested in both preschool and early elementary years. As a result, the HSPV-FT analyses have turned out to be a skeleton of what originally was envisioned.[4]

---

[4] The latest HSPV-FT analysis was conducted by Herb Weisberg and Walt Haney of the Huron Institute. They found that the patterns of data collection in Head Start and Follow Through place severe limitations on the possibilities for estimating combined effects of Head Start and Follow Through. From the limited HSPV-FT data, it does appear that Follow Through maintains increased test scores on skills measured through the end of kindergarten. There were, however, no overall effects apparent on skills measured in later years. Herbert Weisberg and Walt Haney, Longitudinal Evaluation of Head Start Planned Variation and Follow Through (Cambridge, MA: Huron Institute, April 1977), p. i.

Since there has been no chance to apply the pre-test post-test treatment control paradigm to the combined Head Start-Follow Through "treatment," the national Follow Through evaluation has been forced to treat entry into Follow Through as the pre-test time and as the start of the treatment whose effect is to be estimated. From this analytical point of view, the programmatic connection between Head Start and Follow Through has been an important complicating factor. Since more FT children than NFT children attended preschool, this is one of the important characteristics on which treatment and control groups differ. Since preschooling may have given FT children an initial learning boost and consequently raised their pre-test scores, adjustments on the basis of these pre-tests may contain a bias which artifically tends to make FT look bad. This preschool confounding problem, as it came to be called, has generally been handled in two ways. In earlier years children's preschool experience was included as a covariate by which to adjust outcome scores (for example in Emrick, pp. 82-85; and Abt IA, p. IV-6). More recently, investigators have also handled the preschool issue by blocking on preschool experience; that is, separating children with different kinds of preschool experience into separate groups and conducting separate analyses for the different groups (e.g., Abt IVA, pp. 71-80; and Abt IVD, pp. 28-29). In either case, investigators have had to assume for the sake of analysis, that entry into kindergarten is the start of the FT treatment.

End of the Follow Through Treatment and Evaluation. From a programmatic point of view, the Follow Through treatment lasts only through the third grade of elementary school. Occasionally people have advocated extension of Follow Through beyond the third grade to the upper primary grades. For instance, the parent resolution at Denver recommended that the program "should be extended to include the fifth and sixth grades."[5] Despite the occasional suggestions in this vein, this aspect of the temporal issue has been drawn into debates over the program relatively infrequently.

Much more frequently it has fueled arguments over the duration of the evaluation. In part the dispute stems from the fact that the congressional mandate for Follow Through specifies only a very general and long-range goal for the program--to promote the continued development of children to their full potential. The problem of when to evaluate has been recognized by nearly everyone involved in the evaluation effort. For example, in 1973, Emrick and his SRI colleagues wrote:

> Establishing the temporal relevance of evaluation measures to the long-range goals of the program presents still another problem. The general objectives of improving the life chances of poor children suggest that what are needed are sets of measures that validly predict (or diagnose) the long term life chances of such children. In addition, these measures would have to be appropriately sensitive to the effects of specific treatments. The only set of measures that even comes close to meeting these requirements in the current interim data are the achievement scores. But current research has called even these measures into question in terms of their predictive validity and utility. (p. 319)

---

[5] Voices, December 1972, p. 12.

Like so many troublesome issues, the question of when to evaluate was solved mainly out of pragmatic necessity. Early on, evaluators had hoped for longitudinal evaluations of several of the FT cohorts. But after the surge in data collection costs in 1969, the Office of Education decided to limit general data gathering to entry and exit testing. Later it was necessary to restrict the testing even further. As Garry McDaniels wrote,

> By the spring of 1972, base-line data had been collected on [cohort II and cohort III] sites. The cost was so astonishing (ten tons of materials and 2,200 people in the field were needed for a scaled-down effort in the spring of 1972) that the need for testing was reconsidered in terms of cost, effort, and knowledge gained. The eventual decision was to continue studies on the first three cohorts, to study only cities in the fourth, and to collect no data on the fifth. On the first two cohorts, tests would be given only after the children had completed third grade.[6]

Follow-up studies on Follow Through children past third grade--later grades studies as they came to be called--were considered several times. A 1973 review of possibilities cited several "compelling reasons" for a later grades study, and the plan came up again in 1975 and 1976.[7] On all occasions, plans seemed to short circuit mainly because of costs and other practical constraints--like attrition and instrumentation. So the national evaluation has been limited largely to a pre-post design, with post-testing limited mainly to the end of the third grade.

---

[6] McDaniels, "Evaluation Problems," p. 54.

[7] G. Madaus and T. Marx, Report on the Follow Through Later Grades Study (Cambridge, MA: Huron Institute, April 30, 1973). In 1975, several Follow Through sponsors advocated using "language production" test instruments in a later grades study, and a study of "Effects of Follow Through in the later grades" was one of several possible follow-up Follow Through studies investigated by Michael Garet of the Huron Institute. "Follow Through Evaluation Studies in the Next Two Years," Cambridge, MA: Huron Institute, December 1976.

As in so many other cases, the pragmatic solution to the dilemma left some of the Follow Through program participants highly dissatisfied. The discontent on this issue was exemplified by one sponsor's reaction to a draft of an Abt report comparing Follow Through and Non-Follow Through children at the end of third grade:

> An erroneous assumption pervades the entire Abt report. This is the assumption that it is possible to fund a Follow Through model or any other program that will demonstrate immediate, dramatic, and lasting gains in performances on standardized achievement tests. The approach of the Home-School Partnership Model always has been and still is that as the home environment, the school environment, and the environment of the larger community become more supportive of the educational needs of children, such support will pay dividends the increments of which will not occur dramatically in the short run, but will have long-range effects on the child, the child's parents, and the community, and the future offspring of the child. These gains would hardly yield statistically significant results over a brief time span. In our opinion, the Abt report has failed to come to grips with this central point.[8]

The sponsor's plaint is understandable. It does seem unfair to expect a Follow Through model focusing on changing parent and community roles to have as quick an impact on children's academic achievement as a model which directly aims at boosting such achievements.

---

[8] Edward E. Johnson, letter to Dr. Mary M. Kennedy, February 20, 1976.

The worry is that end-of-third grade tests may mismeasure the real and long-term impact of Follow Through in either of two ways. First there is the fade-out fear. Even if the evaluation shows that Follow Through programs have boosted children's spring third grade test scores, it may be that these "effects" will disappear as soon as children leave the program. For example, Wesley Becker and Siegfried Engelmann, sponsors of the Direct Instruction Model, found that follow-up studies of Follow Through children past the third grade are often disappointing. Although the Direct Instruction model has consistently been shown to be one of the most effective Follow Through models in terms of third grade academic achievement test scores, Becker and Engelmann's follow-up studies showed there were "sizeable losses against national norms from third to fifth or third to sixth grade." They concluded, "There is a clear implication that compensatory programs cannot be expected to maintain gains after the programs are stopped, unless all schools are prepared to provide effective teaching for every child in every grade."[9]

The other side of this temporal issue is the worry over what sometimes are called sleeper effects. It may be that an evaluation which focuses on a limited set of third-grade outcomes shows no program effects, but that follow-up studies on other kinds of measures may reveal changes which were simply not apparent at the end of third grade.

---

[9] Wesley C. Becker and Siegfried Engelmann, "The Direct Instruction Model." Eugene, Oregon: University of Oregon, May 14, 1976. pp. 12-13. Emphasis in original. This paper was prepared by Becker and Engelmann as a summary description of the Direct Instruction Model to be included in the OE evaluation volume on Follow Through sponsors.

Both of these possibilities--fade-out and sleeper effects--represent real and important issues in terms of the ultimate goal of aiding in the "continued development of children to their full potential." But at the same time they represent a real dilemma for evaluators and policymakers. After all, the policymaker can hardly be expected to defer evaluation to his or her grandchildren. So it seems reasonable to coordinate evaluation with program--to look for program effects at the end of the program. Particular Follow Through sponsors may indeed have long-term effects, say sleeper effects which grow stronger years after students leave the program. But the prospect of later effects would seem more plausible if some sort of program impact could be shown at the end of the program. What's more, if evaluation were postponed in a search for sleeper effects, then we might find ourselves back where Follow Through started. Having found no apparent effects, say, two years after children leave the program, we might then surmise that the findings suggest a washing out of Follow Through effects after children leave the program. This naturally would bring to mind a new follow through to follow up on old Follow Through.

Standards of Comparison and Outcome Measures

> Perhaps the most important thing Follow Through sponsors have learned is that the power to test in education is the power to control. ...While the testing battery of the national Follow Through experiment represents a realistic acceptance of what is available in standardized tests, it has almost totally failed to capture the potential for educational innovation that Follow Through as a concept promised.
>
> \- David Weikart[1]

The Follow Through evaluation may have been troubled by disputes over when to evaluate, but these disputes seem slight compared to the debate over the standards for evaluation. In terms of the treatment-control paradigm this troublesome question was how to look for program effects, what standards of comparison to use. The story of test instrumentation in Follow Through has three parts. The first goes from the beginning of Follow Through planned variation until the crisis time of summer 1971. The second concerns the selection of a standard test battery in 1971 and 1972, and the third describes the continuing instrument development efforts after 1972.

Early Approaches to Follow Through Instrumentation. Since Follow Through is largely an educational program--in the words of the program rules and regulations, designed to "amplify educational gains"--it was natural that the evaluation effort looked to educational outcome measures. Yet as the Emrick report notes, "It was clear early in the planning and preliminary evaluation activities that for Follow Through to be evaluated as a total program, more must be measured than the participating child's academic progress" (p. 35). Therefore, in the early years of Follow Through, SRI invested much time and effort in trying to develop new outcome measures and in adapting previously existing ones.[2]

---

[1] David Weikart, "Alternative Assessment and National Follow Through," High/Scope Educational Research Foundation Report 1975-76. (Ypsilanti, MI: High/Scope Educational Research Foundation, 1976) p. 4.

During the very first meetings on the planned variation evaluation in the summer of 1968, OE officials and SRI evaluators recognized that no single outcome measure could encompass the range of objectives sought by different sponsors. At the same time they realized that developing altogether new measures would take time.

> A general strategy was adopted provisionally of developing a broad-based pool comprised to test content that reflected objectives more or less common to all sponsors, plus subtests of items which would reflect particular objectives for different sponsors.[3]

The test item pool would, went the description, be like a daisy--with a common center, and petals appropriate to different sponsors' approaches.

SRI decided that the "classroom, not the individual pupil, would be the basic unit of measurement." In this way, the SRI evaluators reasoned, "the goals of the curriculum can be tested at less inconvenience to the child by applying item sampling techniques whereby all classrooms would be subject to a common set of items, but each child would take only a subset of the total battery."[4]

---

[2] Analyses of data from Follow Through parent and teacher questionnaires sometimes employ language suggesting these instruments are outcomes measures; for instance, references to Follow Through program "impact on its teachers" (Abt IIB, p. II-46). Also, Stallings and others sometimes refer to classroom processes as program outcomes. This discussion excludes these types of outcomes not because we think them unimportant, but because attention in the evaluation has been directed mainly at outcomes for children in terms of test scores.

[3] SRI, Draft Administrative History, p. F-11.

The test battery administered in the fall of 1968 consisted of "five booklets of items taken from the nine tests originally selected to represent the core test item pool."[5] Each booklet had three forms and the different forms were randomly assigned to different children in the tested classrooms. During the 1968-69 school year, the SRI evaluators made a number of changes in this battery of tests. Two booklets seemed too easy for first-graders, so these were dropped in favor of subtests of another standardized test. Also, midway in the school year a short form of the Stanford Binet IQ test was administered to a sample of children tested in the fall. Reaction to use of the IQ test "was so adverse that the test data were never analyzed" (p. C-5).

In addition to these efforts the SRI investigators had solicited test items from sponsors in an attempt to develop instruments that could be used to assess objectives unique to the different educational models--the petals in the daisy metaphor. Altogether nine sponsors submitted approximately 1500 items for consideration. After culling these items for duplication and for administrative feasibility and then pretesting candidate items, the SRI investigators grouped roughly 600 of the sponsor-suggested items into two booklets, each with three forms. These sponsor item booklets together with the core instruments resulted in a total of 22 forms of eight tests being administered in the spring of 1969.

---

[4] SRI, Draft Appendix C. Follow Through Pupil Tests, Parent Interviews and Teacher Questionnaires (Menlo Park, CA: SRI, August 1972), p. C-2.

[5] SRI, Draft. Follow Through Pupil Tests, p. C-4. Items in the fall 1968 battery were drawn from the Lee Clark Reading Readiness Test, the Preschool Inventory, the Metropolitan Readiness Test, and the New York University pre-math, pre-science, prepositions, alphabet, numerals and shape names tests. Items from all of these tests were group-administered except for those from the Preschool Inventory (PSI) which were individually administered.

The test battery was further adapted for use in 1969-70 and "to provide a future-time comparison group, a large number of fourth grade rooms from Follow Through schools was also included in the Fall 1969 sample" (p. C-15). In 1969-70 a total of 25 different test booklets was used--all different than those used in the previous school year.

In the third year of planned variation, problems in the proliferation of kinds and forms of tests became evident:

> The apparent advantage of multiple forms of each test began to be outweighed by the obvious disadvantages. Formatting, printing, packing, shipping, receiving, coding and storing posed serious logistic and quality control problems. Multiple forms and the growth of the project made such field procedures as child sampling, following children from one administration of tests to the next on the same test form, tester training and tester record keeping increasingly complex tasks. Therefore, the decision was made to reduce the number of test forms from two forms per grade level to one form for each booklet in the 1970-71 battery. Although this meant that there were now multiple forms of each test given across different grade levels, all tests given at any particular grade level were the same. (p. C-19)

The problems in developing different sets of academic achievement tests to correspond with the planned variation curriculum models were paralleled by difficulties in developing non-cognitive test instruments. In the April 1969 Pajaro Dunes conference, SRI evaluation plans had been criticized for over-reliance on traditional achievement tests. Critics had called for more emphasis on children's noncognitive development.

So in the summer of 1969 SRI formed a committee to look into the issue. The committee identified a number of dimensions as "points of focus" for developing noncognitive test instruments. As had been done in the achievement test development the committee tried to identify specific sponsor objectives in the noncognitive realm. This was done through "both written and verbal interaction with sponsors, ethnic group representatives and consultants." Table 8 shows a list of "sponsor noncognitive objectives" resulting from "such interaction" (p. C-23). SRI considered eight noncognitive instruments as possibly appropriate for evaluating these objectives. Pilot studies in the spring of 1970 revealed the "difficulty of obtaining reliable measures at earlier ages." The SRI summary of the noncognitive test development effort reported that "reliability estimates... are not large enough to warrant inclusion [of the noncognitive instruments] in the Follow Through battery without further development" (p. C-25). In private, at least one SRI staffer put it more starkly: "The money that was spent on the development of noncognitive measures was wasted. Period."[6]

Altogether, the SRI effort in developing new outcome measures appropriate to different sponsors' objectives was a huge disappointment. Among their final 1973 "recommendations concerning evaluations of future education and social action programs," Emrick and his SRI colleagues warned,

---

[6] Elmore, Follow Through: Decisionmaking, p. 307. Elmore's interview with this "prominent SRI staff member" was conducted in July 1972. This verdict on the noncognitive development effort likely reflected the fact that SRI's reporting on the effort was embroiled in considerable controversy (see Krulee, p. 256). The only report on this aspect of SRI's instrument development work was a December 1970 report entitled "Appendix D: Issues in Noncognitive Measurement," which was issued only in draft form and never widely circulated.

Table 8

SPONSOR INTERESTS ACCORDING TO MEASUREMENT AREAS

	Sponsor								
Desired Outcome of Model	**Far West (02)**	**U. Arizona (03)**	**Bank Street (05)**	**U. Oregon (07)**	**U. Kansas (08)**	**High/Scope (09)**	**U. Florida (10)**	**EDC (11)**	**U. Pittsburgh (12)**
I. ATTITUDES TOWARD SCHOOL									
A. Positive Attitudes and Behaviors		X		X	X		X		
B. Enthusiasm				X	X			X	
II. TASK ORIENTATION SKILLS									
A. Attention to Task		X		X	X	X			X
B. Resist Distractions	X	X	X			X	X		
C. Task Persistence		X		X	X	X			
D. Desire for Mastery	X						X		
III. CURIOSITY (Explores Environment)	X	X	X					X	
IV. AUTONOMY (Independence)									
A. Autonomous Learner (choose, plan)	X	X	X			X			X
B. Self-Management			X			X		X	
C. Initiates Activities		X	X				X	X	
D. Problem Solving	X					X			
V. SELF-ESTEEM									
A. High Self-Esteem				X	X		X		
B. Positive Self-Concept	X		X			X	X	X	
VI. SCHOOL FEARFULNESS									
A. Low School Fearfulness			X					X	
B. Open and Receptive			X			X		X	
VII. LOCUS OF CONTROL									
A. Increase Sense of Personal Control Over Behavior Consequences							X	X	
B. Challenges Ideas of Others								X	
C. Expect Success		X							

Source: SRI, Draft Appendix C. Follow Through Pupil Tests, Parent Interviews, and Teacher Questionnaires, p. C-24. This table should not be considered a completely unambiguous identification of sponsor "interests" in the noncognitive realm. Krulee notes: "Some friction could easily and did develop between SRI staff members and sponsors...over the program of noncognitive testing. To some extent SRI expected and needed help from sponsors in defining what ought to be done in this area and in the development of hypotheses. It was not always possible to obtain this help." Krulee, Organizational Analysis, p. 256.

> The temptation in undertaking an evaluation is to assume that procedures can be rapidly developed or that ways to deal with problems of measurement and analysis will emerge. There is, however, great danger in basing one's plans on the assumption that such developments will occur...While the importance of measures in the noncognitive domain was clearly recognized, neither limitations in existing instruments... nor the difficulties inherent in effecting new developments were equally clearly recognized, and certainly not on a schedule that could be useful to the evaluation. Evaluation planning, in short, should be limited to the state of the art as known or best estimated at the time the evaluation begins so that unrealistic expectations about new developments can be avoided. (p. 323)

Long before this, however, dissatisfaction with SRI instrument development efforts built up for a variety of reasons. The SRI strategy of using the classroom instead of the individual as the "unit of measurement" may have made sense in terms of exploratory instrument development, but it made longitudinal evaluation over more than one school year nearly impossible. Also, the nonstandardized testing raised problems when Richard Fairley, OE Director of Compensatory Education, visited SRI in February 1971. Fairley wanted evaluation results expressed in grade equivalent terms, but the test battery included "no intact measure with acceptable norms from which to deduce effects in grade equivalent terms."[7]

Moreover, how to combine disparate items into outcome measures was not at all clear. This ambiguity led to a decision in 1970-71 to classify all items "as belonging to one or more of the six basic Dependent Variable classifications" that SRI had established.[8] The six categories were:

---

[7] SRI, Draft Administrative History, p. F-28.

[8] SRI, Draft Follow Through Pupil Tests, p. C-19. In addition to these categorical distinctions, SRI applied item selection criteria pertaining to item difficulty, discrimination and subtest reliability.

WRAT; Quantitative; Cognitive Processes; Reading; Language; and Affect. The sum of items classified into each of these categories formed the basic outcome measures for the Emrick report. In addition the Emrick study employed an overall achievement outcome measure--simply the "raw score sum of correct responses to all cognitive test battery items" (p. 27). This practice led to a peculiar result, one that was quite contrary to the spirit of creating sponsor-specific outcome measures. The eight were separately named, but in fact contained many of the same items. For example, "The WRAT achievement variable...overlaps with both the total achievement variable, and the quantitative, reading and language variables" (p. 29).

Although using the same SRI test item information, Robert Soar did not employ the same outcome measures as Emrick. Instead of grouping the item data into the SRI categories, Soar used factor analysis to group 1970-71 data into three to five outcome measures for each grade level.[9] So the early experimentation with a range of new test items had an unforeseen consequence. The two major national evaluation studies analyzing spring 1971 outcomes, though based on much of the same pupil test data, employed different outcome measures.

Selection of a Standard Test Battery. The issue of instrumentation for the Follow Through evaluation came to a head in the summer of 1971. Upon reviewing materials submitted by SRI in September 1971, OE officials expressed concern over continuity of testing, the timetable for instrument development work and the rationale behind past testing. The reviewers felt that

---

[9] Soar, Classroom Process, pp. 34-35.

"the present battery lacked adequate instrumentation in noncognitive areas..." and that "the present battery was also very limited in its measurement of cognitive areas."[1] As a result, they decided that the Office of Education would take over responsibility for selecting instruments to be used in the spring 1972 test battery.

So in late 1971 and early 1972, the new chief of Follow Through evaluation, Garry McDaniels, together with consultants and a special instrument selection panel took on the task of choosing new test instruments. The battery they chose was to be administered as each cohort stream of children completed Follow Through. It consisted of the MAT, the Ravens, the Coopersmith, and the IARS.[2]

Metropolitan Achievement Test. The MAT is a standardized achievement test published by Harcourt, Brace, Jovanovich, Inc. Four versions of the instrument have been used in the Follow Through evaluation: the MAT Primer at the kindergarten level, the MAT Primary I at first grade, the Primary II at second grade, and the Elementary version at third grade. The publisher's description of the subtests and contents of each of these subtests is given in Table 9. Intermediate testing at first and second grade was done for some sites in every cohort but was heaviest for cohort III. For cohorts I,

---

[1] Judy Crooks and Donald Burns, "Consolidated Review of September 17 Deliverables" memorandum to Mr. Fred Bresnick, October 6, 1971.

[2] The best summary of the full span of Follow Through testing is given in Anne R. Wright, Summary Report on Follow Through Testing Program, 1969-1975 (Menlo Park, CA: SRI, September 1976). For more specific information on which tests were administered for which cohorts at which time points, see SRI, Follow Through Evaluation Data Bank File Description All Years Beginning 1972-1973. Part I; Revised January 1975, pp. I-4-I-18.

Table 9: Metropolitan Achievement Tests Used in the National Follow Through Evaluation: By Version, Subtest and Content

PRIMER

Listening for Sounds: 39 items measure pupils' knowledge of beginning and ending sounds and sound-letter relationships. Twenty-two of these items are based on pictures; eight items are based on letters; nine items use single words.

Reading: 33 items measure pupils' beginning reading skills. Eleven items require pupils to identify letters; fourteen items require pupils, when given a picture of some common object, to select from four words the one word that describes the picture; five items require pupils to select one of three easy sentences which best describes a picture.

Numbers: 34 items measure pupils' understanding of basic mathematical principles and relationships. Twenty items cover counting, measurement, numerical recognition, etc.; fourteen items measure pupils' ability to add and subtract one-digit numbers.

PRIMARY I

Word Knowledge: 35 items measure extent of pupils: reading vocabulary. Pupils are given a picture of some common object and must select from four words the one word that describes the picture. Words are generally from primary level readers.

Word Analysis: 40 items measure pupils: knowledge of sound-letter relationships or skill in decoding. Pupils must identify a dictated word from among several words with similar configurations and sound patterns.

Reading: 42 items measure pupils: comprehension of written material. Thirteen items require pupils to select one of three easy sentences which best describes a picture. Nineteen items require pupils to read simple paragraphs and answer questions about what they have read.

Mathematics - Part A:Concepts: 35 items measure pupils: understanding of basic mathematical principles and relationships. Items cover counting, place value, sets, measurement, etc.

Part B: Computation: 27 items measure pupils: ability to add and subtract one- and two-digit numbers with no regrouping.

PRIMARY II

Word Knowledge: 40 items measure extent of pupils' reading vocabulary. Seventeen items are in the word-picture association format. Twenty-three items require pupils to identify a synonym, antonym, or classification for a given word.

Word Analysis: 35 items measure pupils' knowledge of sound-letter relationships or skill in decoding. Pupils must identify a dictated word from several printed words which have similar configurations or sound patterns.

Reading: 44 items measure pupils' comprehension of written material. Thirteen items require pupils to select one of three sentences which best describes a given picture. Thirty-one items require pupils to read a paragraph and answer questions about what they have read.

Spelling: 30 items measure pupils' ability to spell commonly used words. The test uses the familiar format in which the teacher reads a word and pupils write the correct spelling.

Mathematics: Computation: 33 items measure pupils' ability to compute. Items cover addition of one- and two-digit numbers, three addends and missing addends, subtraction of one-, two-, and three-digit numbers, and a few simple multiplications. Some regrouping and some horizontal notation are introduced.

Mathematics: Concepts: 40 items measure pupils: understanding of basic mathematical principles. Place value, measurement, laws and properties of number systems, arrays, sets, inequalities, etc. are covered.

Mathematics: Problem Solving: 35 items measure pupils' ability to apply knowledge in solving numerical problems. About one half of the items are dictated to pupils, whereas pupils read the remaining problems to themselves. Problems cover addition, subtraction, multiplication, and division processes, rate. multiple-step problems, and use of number sentences.

ELEMENTARY

Word Knowledge: 50 items measure extent of pupils' reading vocabulary. Items require pupils to identify synonyms, antonyms, or word classification. Items range from primary level to junior high level in difficulty.

Reading: 45 items measure pupils' ability to comprehend written material. Pupils read a paragraph and then answer questions about it. Items cover comprehending literal meanings of passages, drawing inferences from the material, identifying main ideas, and determining word meanings from context.

Language: 50 items measure pupils' knowledge of basic conventions in standard written English. Fifteen items require pupils to identify whether given sets of words are "telling" sentences, "asking" sentences or not sentences at all. Thirty-five items require pupils to identify errors in punctuation, capitalization, or usage in written material.

Spelling: 40 items measure pupils' ability to spell commonly used words. The test uses the familiar format in which the teacher dictates a word and pupils write the correct spelling.

Mathematics: Computation: 40 items measure pupils ability to compute. Items include 10 addition, 9 subtraction, 7 multiplication, and 7 division examples with integers (some requiring regrouping and some in horizontal notation), and 7 items on decimals and fractions.

Mathematics: Concepts: 40 items measure pupils' understanding of important mathematical principles and relationships. Concepts covered include laws and properties of number systems, measurement and geometry, place value, sets, etc.

Mathematics: Problem Solving: 35 items measure pupils' ability to apply knowledge in solving numerical problems. Items cover application of ASMD processes to everyday problems, chart reading, and use of number sentences.

Source: W. N. Durost, et al., Metropolitan Achievement Test Tester's Directions, New York: Harcourt Brace, 1971.

II and III, the MAT Elementary test was given at the end of third grade as pupils finished the Follow Through program. Cohort IV was tested only through the second grade.

The MAT was selected by the review panel because,

> It was felt that achievement should be assessed by an instrument with adequate psychometric properties to measure skills in language acquisition and use, reading, and numerical facility. Of the standardized achievement tests available for review, the MAT seemed to provide adequate coverage of skills and to contain fewer obviously biased items (cultural and ethnic bias). It was also deemed desirable for comparison purposes to have a single instrument which could be used at several grade levels. the MAT, with its [different] versions, meets this requirement.[3]

Ravens Coloured Progressive Matrices. The instrument selection panel chose the Ravens because of the possibility that the MAT might provide an unfair advantage to sponsors who focused heavily on specific skill acquisition in contrast to those who stressed more general cognitive development. The 1972 review panel commented:

> Some sponsors felt that an achievement test such as the MAT contains such a narrow academic focus that it would not do justice to their programs. This opinion was expressed by sponsors who feel that their programs focus on flexible problem solving approaches rather than academic content. There was also some feeling that behavioristic programs may teach the content of such sub-tests as letter recognition, beginning sounds, counting, etc., thereby, giving children in these programs a testing advantage based on practice. *Ravens Coloured Progressive Matrices* test was included in the battery as a task which would provide all children with an opportunity to demonstrate their problem solving skills. It was agreed that this was a task which could not be

---

[3] OE, *Second Report of the Follow Through Review Panel*, p. 23.

> taught or for which children could not be coached. Therefore, no sponsor or group of sponsors should find the test biased against any type of program. However, we do not believe that this test is free of cultural bias, but this will not constitute a problem as long as no comparisons are made across ethnic groups.[4]

The Ravens consists of a series of diagrams or pictures from which a piece is missing. Children are asked to select the appropriate missing piece from among a set of alternatives. The instrument was originally designed to be administered individually. However, SRI with the cooperation of test author John Raven, revised the instrument to be administered in the group mode. The SRI version omits the two final items for each of the three sets of patterns contained in the original version, and so has only 27 items.[5]

Intellectual Achievement Responsibility Scale (IARS). The IARS is a non-cognitive instrument designed to assess the extent to which children attribute responsibility for their own successes and failures either to themselves or to external influences. The Follow Through review panel selected the IARS Scale as a "measure of 'locus-of-control' or feelings of competence about one's ability to influence important events in his life."[6]

Coopersmith Self-Esteem Inventory. The Coopersmith was designed to measure children's feelings of self-esteem both in general and in a school setting. The test is group-administered and consists of a series of statements--such as, "kids pick on me very often"--after which children are asked to say if the sentence says something that is "like me" or "not like me." In early years the form of the Coopersmith used in the Follow Through evaluation was a special shortened 32-item version, but from 1973 onward, the standard 58-item Coopersmith was used. Since 1972 the Coopersmith has been used only at the end of third grade.[7]

---

[4] OE, Second Report of the Follow Through Review Panel, p. 23. Emphasis in the original.

[5] Wright, Summary Report, p. 54.

[6] OE, Second Report, p. 24.

The 1972 review panel recommended the two noncognitive tests only with reservation:

> Many sponsors, parent groups, educators and psychologists have been concerned about the development of non-cognitive aspects of children's personality during Follow Through. Questions about ethnic identity, sense of control over the environment, self-confidence, self-esteem, self-concept of ability, school or test anxiety, and achievement motivation have been raised. Even though the panel feels that there are no really adequate measures in these areas at this time, it recommends the use of two of the best available measures at this time in order to attempt to answer some questions about a child's socio-emotional growth.[8]

The panel also expressed doubts about the cognitive tests. The "general feeling" among panel members was that the use of only the MAT and the Ravens "leave unanswered many questions about children's learning ability, and the manner in which children acquire, process and utilize knowledge" (p. 24). The panel concluded:

---

[7] Wright, Summary Report, p. 55.

[8] OE, Second Report, p. 24.

> the cultural biases and other inadequacies of standardized achievement tests make it imperative for the impact of Follow Through on other aspects of child development to be assessed. However, the inadequacy of currently available non-cognitive measures suggests that new approaches need to be considered. After considering the magnitude of the development task, however, the panel recommends that test development should not be a part of the Follow Through evaluation effort.[9]

Despite these reservations, a standard set of outcome measures had been selected for assessing Follow Through treatment effects at the end of third grade: the Metropolitan Achievement Test, the Ravens Coloured Progressive Matrices, the Intellectual Achievement Responsibility Scale and the Coopersmith Self-Esteem Inventory.

In years prior to the third grade evaluation analyses, the Stallings studies and the Abt I and II reports of kindergarten and first grade outcomes relied also on outcome measures previously selected and administered by SRI. In the academic achievement realm these were:

> Wide Range Achievement Test (WRAT). This instrument is an individually administered achievement test designed to measure skills in the areas of reading, spelling and arithmetic. The form of the WRAT administered in FT was adapted from a standard one so norms for the standard instrument were not applicable to the FT test score data.
>
> Peabody Picture Vocabulary Test (PPVT). This individually-administered test was designed to measure a child's verbal intelligence by assessing the number of words the child knows as he or she hears them. As words are spoken the child is asked to point to pictures corresponding to spoken words. The form of the test used in FT was adapted from the standard test by modifying some picture to reflect Black rather than White ethnic characteristics.[1]

---

[9] OE, Second Report, pp. 24-25.

[1] Scoring procedures and details of administration of these two tests are given in Wright, Summary Report, pp. 19-33 and 46 respectively.

In the noncognitive domain two other instruments were also treated as outcome measures in the early Abt reports:

> Gumpgookies. This test is designed to measure young children's motivation to achieve in school. Each of the test's items involves a semi-structured story about two imaginary figures called Gumpgookies. After an examiner reads the story describing the action of the two Gumpgookies, the child is asked to choose the Gumpgookie in the story which is most like him or her, for instance the Gumpgookie which is forgetful or the one that is not.
>
> Locus-of-Control. This instrument is designed to assess the child's feelings of the extent to which he or she is responsible for things happening in his or her life in comparison to the child's attribution of responsibility to other people or forces. The events described are limited to school-related situations and are divided into positive and negative happenings. This distinction yields two subscores from the instrument--Locus-of-control-positive and Locus-of-control-negative. The version of the Locus-of-Control used in FT has 20 items, one less than the standard instrument.[2]

In addition to test score data, the Emrick, the Abt I and II and the Soar evaluation studies also treated absence--the number of days a child is absent during a school year--as an outcome measure. The Abt III and Abt IV reports dropped absence as an outcome measure and the Abt IV report explained,

> The number of days a child missed school throughout the kindergarten year was analyzed as an indirect measure of a general capacity for a sponsor to attract pupils consistently. While children who enjoy school are more apt to attend, this interpretation cannot be fully accepted without qualifications concerning other reasonable determinants of absence such as sickness, parental factors, or environmental phenomena operating within the community.[3]

---

[2] The Locus-of-Control test uses four different sets of pictures in order to prevent sex and racial biases; the four versions are for Black Boys, Black Girls, White Boys and White Girls. Wright, Summary Report, p. 47. See p. 48 of the same report for more information on the use of Gumpgookies in Follow Through.

[3] Abt IVA, p. 39.

In summary, numerous test instruments have been employed at one time or another during the Follow Through evaluation, but the final Follow Through third grade outcome national evaluation is based on only four; the Metropolitan Achievement Test, the Ravens' Coloured Progressive Matrices, the Intellectual Achievement Responsibility Scale, and the Coopersmith Self-Esteem Inventory. No one has been completely satisfied with this battery as the sole standard of comparison for the national Follow Through evaluation. Even the panel which selected the battery expressed doubt about it. In the years since 1972 these doubts and dissatisfactions have led to numerous attempts to find new ways to judge the effectiveness of Follow Through.

Continuing Instrument Development Efforts. When the standard test battery was chosen in 1972, it was recognized that some sponsors whose goals did not mesh well with the standard instruments might be put at a disadvantage in the national evaluation. Therefore, in funding evaluation efforts by sponsors, McDaniels gave high priority to sponsor objectives not included in the national evaluation effort. The high priority rating did not extend, however, to the development of new instruments to assess these objectives. Since the time remaining to document program impact was extremely limited, and because instrument development had proved so troublesome in the past, McDaniels and the evaluation review panel rated instrument development "lowest on the list of activities for sponsors to pursue."[4]

---

[4] Elmore, Follow Through: Decision-Making, p. 308. Second Report of the Follow Through Evaluation Review Panel, p. 20.

For a time, instrument development work in Follow Through seemed to wane. But despite the "lowest priority" rating for the efforts, phoenix-like they seemed to reappear--both within the national evaluation and in sponsors' evaluation activities. At the national level, the instrument development work took the form of reanalyzing items from the standard battery in the hope of uncovering dimensions of program impact which were most apparent in the overall instruments. In 1974-75, one staff member at the Huron Institute attempted this kind of instrument development.[5] Among other things he tried to identify subsets of items in the standard MAT which might be more sensitive to Follow Through program effects than was the overall instrument. In a draft report in July 1975, he concluded: "[T]his kind of secondary item analysis using items on standardized tests which have already been through the psychometric wringer is not likely to be fruitful."[6]

Like the SRI effort, this work resulted in neither new instrument nor final report. Yet despite the small return on previous efforts to develop new instruments for the Follow Through evaluation, the activitiy recurred even as the national evaluation drew to a close. The main Abt analyses in 1975-76 were based upon the four standard instruments selected on 1972. Yet one part of the Abt evaluation analyses in 1976 consisted of an effort much like the 1974-75 effort at Huron--an attempt to develop new outcome measures--what the Abt analysts called quasi-objective referenced instruments--out of subsets of the MAT items.[7] Like each of the previous test development efforts, this one too dropped by the wayside and never was described in any final reports.

---

[5] F. Mosher, "Progress Report on the Instrumentation Study," Huron Institute, June 1, 1974.

[6] F. Mosher, Follow Through Instrumentation Study. Draft version. July 1975, p. 73. Emphasis in the original.

From an outside distance it seemed to be a cycle of the absurd: instrument development efforts which yielded no apparent payoff for the evaluation; followed by relegation of the endeavor to lowest priority; followed by more and yet more seemingly fruitless instrument development work. Yet close up, in context of the history of the Follow Through program and the national evaluation, the recurring pattern of attempts to develop new instruments becomes more understandable. Perhaps the main motivation for the repeated attempts to find new outcome measures was the widespread feeling that the MAT, the Ravens, the IARS and the Coopersmith span only a very narrow range of the realm of aspirations for the program.

Clearly it is this feeling that has prompted continuing efforts by Follow Through sponsors to develop new test instruments more appropriate to their own goals. Indeed the first of the annual sponsor conferences, held in October 1971, was originally called a "Working conference on evaluation of Follow Through objectives not measured by academic tests." Working groups at the conference dealt with a variety of measurement issues; specifically, unobtrusive/nonreactive measures, classroom observations, attitudinal measures, and learning-to-learn measures. Despite McDaniels' relegation of instrument development to lowest priority for sponsor evaluation activities, sponsors returned to instrument issues in their second annual conference in October 1972, with working groups on unobtrusive/nonreactive measures and classroom observations. In later years the sponsors have focused even more attention on instrument development. Annual sponsor conferences in 1973, 1974, and 1975 devoted continuing attention to the topics. For example, the third annual conference report listed some 60 instruments developed and/or used by sponsors to "assess special aspects of their model." And in 1975, more than half of the fifth annual conference report was devoted to similar descriptions of the new instruments.[8]

---

[7] Abt Associates, "Follow Through Design Conference, July 19-20, 1976 Conference Handout," Cambridge, MA: Abt Associates, Attachment R. The Abt 1976 effort was somewhat different than Mosher's because it draw on a much wider data set than did Mosher.

Sponsors' continuing efforts to develop new test instruments have taken several forms. Bank Street staff, for example, have been developing a set of instruments for observing patterns in children's communication and behavior in educational settings.[9] The High/Scope Educational Research Foundation has been working on an instrument to analyze children's abilities to utilize written language as an expression of conceptual ability.[1] Follow Through staff at the Education Development Center have been experimenting with a system for analyzing daily and weekly logs kept by teachers and parents of Follow Through children.[2]

---

[8] High/Scope, reports of annual sponsor conferences, see bibliography.

[9] The Bank Street system was developed by Garda W. Bowman and Rochelle Selbert Mayer and is called BRACE (Behavioral Ratings and Analysis of Communication in Education).

[1] David Weikart, "Alternative Assessment and National Follow Through," in High/Scope Educational Research Foundation Report 1975-76, p. 5. The High/Scope instrument is called PLAT (Productive Language Assessment Tasks).

[2] David J. Marion, Gene Nameche, and Jacqueline Rogers, Phase One 1975-76 Documentation Study. Newton, MA: Education Development Center, December 1976.

Taking an even less structured approach the staff of the North Dakota New School Approach have concentrated most of their evaluation efforts on teacher, child and parent interviews.[3]

Though Follow Through sponsors' instrument development efforts have been exceedingly diverse, sponsors seem fairly united in several overall principles. First, they want to develop instruments which will allow assessment of a wider range of skills and abilities than are tested in the Follow Through evaluation battery. Second, they want to assess children's performance in situations more natural than the formal setting of group-administered standardized tests. Third, they generally want to develop instruments not only useful for judging program effectiveness, but also usable by participants themselves for improving the program.

These concerns are by no means unique to Follow Through. They are part of a widespread and growing debate over the role of standardized testing in American education. The movement against standardized testing began in the 1960's with concern over IQ testing, and widespread worries that standardized tests of ability discriminate against minority children. Since then the movement has grown to encompass widespread re-examination of all standardized testing. Numerous professional groups have called for systematic study of the role of standardized testing in American education and several have advocated a moratorium on standardized testing.[4] In 1976, Terry Herndon, the executive director of the National Education Association, asserted straight out that "standardized testing must go."[5] In April 1977, Congressman Michael Harrington introduced into Congress the Testing Reform Act of 1977 designed "to promote the rights of individuals who take standardized educational tests; to make such individuals, educational institutions, and other persons more aware of the uncertain reliability of such tests and the problems of using such tests; and for other purposes."[6] In May 1977, the lead article in the New York Times spring survey of education was headlined "Controversy Over Testing Flares Again," and was devoted almost entirely to the growing debate over standardized testing.[7]

---

[3] University of North Dakota Center for Teaching and Learning, "The New School Approach" paper prepared for OE evaluation volume of sponsor chapters, 1976.

---

[4] A review of several dozen calls for such action is given in Robert L. Williams, Doris Mosby, and Verlaine Hinson, "Critical Issues in Achievement Testing of Children from Diverse Ethnic Backgrounds," paper presented at the U.S. Office of Education Invitational Conference on Achievement Testing of Disadvantaged and Minority Students for Educational Program Evaluation, Washington, D.C. May, 1976.

[5] Herndon's call for doing away with standardized tests was based on his belief that the interpretations of tests are usually based on "several fallacious assumptions about education," that standardized testing fosters "big brotherism," that it "encourages conformity at the expense of creativity," and that standardized tests simply do not do what they are claimed to do" (that is, "improve the allocation of resources," "provide data to help parents evaluate their schools and make decisions about educational needs," and "test results can be used to evaluate teachers performance"). Terry Herndon, "Standardized Tests: Are They Worth the Cost?," The Massachusetts Teacher, April 1976, pp. 22-24.

[6] Congressional Record May 2, 1977. Harrington's bill was referred to three subcommittees of the House Committee on Education and Labor. As of August, no action had been taken on the proposal.

[7] New York Times, May 1, 1977.

So the controversy over instrumentation for the Follow Through evaluation is merely a reflection of a wider and continuing national debate over educational testing.[8] The debate has roused passions in part because it is entangled with highly controversial issues regarding the heritability of intelligence, the growth of federal control over education, and the monopoly powers of the major test publishers. At root, however, the controversy seems to center around a very real tension between the varied goals of American education and the desire to assess alternative ways of investing limited resources in the education of American children. Those who emphasize the diversity of the educational endeavor tend to stress that culturally diverse children in communities with different social standards should not, indeed cannot, be measured with the same yardstick. Those who stress the need for improving investment of scarce resources tend to emphasize the potential of systematic evaluation using common standards of comparison to help people select between alternative means for achieving common educational goals.

This tension has stretched the Follow Through evaluation in several ways. In the early years, the SRI investigators attempted to develop different test instruments which would respond to varied goals of different sponsors. When the Office of Education took over responsibility for selecting tests for the evaluation, instruments aimed at assessing noncognitive aspects of children's personalities were included in the test battery even though the instrument selection panel felt that there were no really adequate measures in the noncognitive area. And despite the low priority rating given to instrument development work by OE's Follow Through Branch in 1972, considerable time and energy have been invested by Follow Through sponsors in efforts to develop techniques for assessing goals not encompassed in the national evaluation battery.

---

[8] Several Follow Through sponsors, in particular Vito Perrone of the University of North Dakota, have become prominent in the movement against standardized testing. For a recent collection of papers on standardized testing see Paul L. Houts (ed.), The Myth of Measurability (New York: Hart Publishing Co., 1977).

The degree of effort to develop new instruments has varied widely among sponsors, and mainly in predictable patterns. Those who adhere more to traditional goals of schooling and aim mainly at readjusting the means of schooling are little troubled by standardized testing and the national evaluation test battery. For example, Wesley Becker and Siegfried Engelmann of the Direct Instruction Model maintain as one of their basic assumptions that "the learning of basic skills is essential to intelligent behavior and should be the main focus of a compensatory education program."[9] Not surprisingly their evaluations rely almost exclusively on standardized test instruments.

More toward the reform end of the spectrum of sponsors is the EDC Open Education approach. Sponsors of the EDC approach who hold as their "core purpose and goal, the full humanization of the teaching/learning process," maintain flatly that "the standardized tests which were used in the national

---

[9] Wesley C. Becker and Siegfried Engelmann, "The Direct Instruction Model," paper prepared for the U.S. Office of Education volume of Follow Through sponsor chapters, May 14, 1976.

Follow Through evaluation are not appropriate to the process oriented goals of the EDC model."[1]

Evaluation based on standardized tests of course has little power to illuminate such conflicting views on the goals of education. More likely it might help us judge between alternative means of reaching the particular goals implicit in such tests. The resolution of more fundamental questions regarding the priorities given to such goals rightly must be left to personal choice and political processes.

Attrition

At a 1976 presentation on the problem of attrition in the Follow Through program one observer quipped, "Attrition is like a social disease. It's a problem we all have, but nobody talks about."[2] As with a social disease, the attrition problem is not very consistently defined, nor is there much agreement about how to solve it. But at least in Follow Through, evaluators have been probing the problem.

This account of the attrition problem in the Follow Through evaluation is divided into four parts. The first simply describes the problem. The second recounts early handling of it. The third summarizes how it was treated during the first phase of evaluation at Abt, and the fourth relates how it was handled during the final phase at Abt, with the curious result that while the attrition problem *looked worse*, people seemed to *feel better* about it.

---

[1] Education Development Center, "EDC Open Education Follow Through Program," paper prepared for the U.S. Office of Education volume of Follow Through sponsor chapters, August 1976.

[2] The anecdote was recounted to me by Elizabeth Proper of Abt Associates, who presented a paper on the attrition paper at the 1976 American Educational Research Association conference in San Francisco.

Defining the Follow Through Attrition Problem. The attrition problem, simply put, is what should be done about treatment subjects for whom we don't have data at the end of the treatment period? The problem is, of course, more complicated than this question suggests because post-test data can be missing for a variety of reasons. This has led to a variety of efforts to define the attrition problem in the Follow Through evaluation into several components.

In a 1972 discussion, for example, Garry McDaniels suggested that "two dominant types of attrition are census attrition and test attrition."

> A census absence means that the child has left either the treatment (sponsored) or comparison classrooms and has not appeared again. This attrition is documented in the census taken of the Follow Through population each fall and spring...Test attrition refers to the child who is absent at the time of testing for all or part of the battery while remaining a part of the target group according to the census.[3]

Four years later, Abt analysts working on the attrition problem defined the idea differently.

> The phenomenon of attrition may be defined in a number of ways. The definition developed for use in the Follow Through evaluation is a good deal more comprehensive than most, and consists of three major categories: Policy, Program and Sample.

---

[3] Garry McDaniels, "Attrition as an Analytical and Programmatic Problem for Head Start Planned Variation and Follow Through." Paper presented at the 1972 APA meetings, August 1972, pp. 3, 4.

> Policy attrition is defined as the administrative dropping of a unit (at any level) after the initial implementation of a program. Within the Follow Through evaluation, this might mean the permanent cessation of data collection in an entire site, or within a single school in that site. Policy attrition may even occur at the child level if for instance the policy of a given school is to return all children whose academic performance has improved to a "regular" classroom in order to provide space for the more needy child.
>
> Program attrition defines the types of subject behaviors that are generally thought of when attrition is mentioned. This includes the loss of subjects due to mobility, dislike of the treatment, illness or death, etc....
>
> Sample attrition occurs within the office of the evaluator and/or data collector. Included in this category are subjects who are omitted from the analytic sample for any number of data deficiency reasons including an invalid test, a large amount of missing data, or inadequate cell size.[4]

The sorts of distinctions made by Abt and McDaniels have, however, generally fallen by the wayside in the heat of evaluation. The Abt analysts recount:

> While we have conceptualized attrition as being of three major types, this year we have not distinguished between different types of attrition in our analyses. This was done because our primary concern has been to ascertain whether the analytic sample was representative of the sample which was present at the beginning of the program. In order to answer this question it was necessary to collapse across our attrition definitions.[5]

---

[4] R.F. St. Pierre and E.C. Proper, "Attrition: Identification and Exploration in the National Follow Through Evaluation." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976, p. 4.

[5] St. Pierre and Proper, "Attrition," p. 4.

Are the treatment and control groups tested at post-time the same as, or at least essentially representative of, the treatment and control groups at pre-test time? This is the form of the attrition problem which has usually been in question in the Follow Through evaluation.[6] Even without considering the different causes of attrition--that is, the different reasons why an evaluator might not have post-test scores for some of the children who started out in the experiment--the attrition problem has proved susceptible to no easy solution. It has, however, been subject to numerous forms of attack.

Early Handing of the Attrition Issue. The attrition problem in Follow Through is roughly analogous to the problems of non-response in survey research and missing data in experimental research. In survey research, the preferred manner of handling the problem is to resurvey a sample of initial nonresponders to make certain that nonresponse is not systematically related to particular answers respondents give.[7] In experimental research, it is usually advocated that a researcher try to determine why particular data are missing before deciding how to handle the issue.[8] The trouble in Follow Through is that the evaluation is of such size and geographic scope that neither of these solutions is practically possible. As a result Follow Through evaluators have been forced to use less direct methods.

---

[6] Note that this discussion of attrition focuses on the issue of internal, rather than external, validity of findings.

[7] For example, C.H. Backstrom and Gerald D. Hursh, Survey Research (Evanston, IL: Northwestern University Press, 1963), pp. 50-59.

[8] For example, G.W. Snedecor and W.G. Cochran, Statistical Methods, 6th ed. (Ames, IA: Iowa University Press, 1967), pp. 317-20.

The most indirect attack on the issue came in the early evaluation studies which used classroom as a unit-of-analysis. This strategy discounted the need for post-test scores on all students by assuming that every student in a class is representative of the class. On this assumption, it is necessary only to have post-test scores on a sample of children in a classroom and not on everyone in each class. This was the strategy employed in both the Soar and the 1972-73 Stallings reports in the analyses of the relationship between classroom observations and pupil achievement. For example, Soar eliminated from his sample students who scored high on the pre-test and so had little room to gain on the post-test (p. 37). He acknowledged this sort of purposeful elimination of some post-test scores--what the Abt analysts call sample attrition--as a major limitation in his findings (pp. 168, 202), but does not explicitly state the minimum number or proportion of pupil scores per classroom which he required in order to calculate a usable class average score.[9]

In her 1972-73 report, Stallings is more explicit on the minimum criteria for calculation of class average scores. "In order for a classroom to be included in these analyses, at least 20% of the children in the classroom had to have baseline test data."[1] Given this criterion, Stallings gives no further attention to the problem of student attrition. There does, however, seem to be a problem of attrition at the classroom level, that is, whole classes being dropped from her analysis. For example, in the correlation and multiple regression analyses of the relationship between classroom observation variables and outcome measures, different sets of classrooms seem to have been used without any explanation of the difference (166 classrooms were used in the correlational analyses and 162 in the regression analyses; see pp. 275-316).

---

[9] For example, he states that "eliminating pupils with high pre-test scores reduced some classroom N's below a level which permitted analysis, however, so the number of classrooms decreased further" (p. 29). In a personal communication, Soar explained that seven was his minimum number of students with data required for retaining a classroom in his analysis.

[1] Stallings, 1972-73, p. 233. In her 1971-72 report, which used the child as the UA, Stallings did not discuss the attrition issue per se, but instead treated the broader issue of missing data. She tested the "effects" of missing data in regression analyses, including dummy variables for missing data, and concluded that "These patterns suggest that the effect of ignoring cases with missing data is different at different sites" (p. 533).

Using the classroom as the basic unit-of-analysis was natural in the classroom observation studies conducted by Soar and Stallings since most of the data they analyzed were drawn from observing classrooms as units. The use of the classroom as the unit-of-analysis in Emrick's study of program impact on pupil outcomes was more problematic. As previously noted, part of Emrick's justification for analyzing classroom average scores was the attrition problem.

> [I]f program attrition is randomly distributed across subgroups (FT/NFT) within projects, aggregation to classrooms will tend to reduce the apparent impact of this attrition; i.e., even though some or most pupil data within a classroom are lost to attrition, under the assumption that the attrition is random, an unbiased mean can still be estimated for the classroom from the remaining pupil data.[2]

Emrick examined the assumption of unbiased attrition by comparing classroom mean scores for the traced subset of pupils who had both pre- and post-test data (what Emrick called the identic subset) with the overall classroom mean scores derived from all pupils who took the tests at either pre- or post-test time. He found that the classroom mean pre-test scores for the two groups differed very little, but that for post-test scores, the classroom mean scores were generally higher for the traced sample than for all pupils taking the post-test.

---

[2] Emrick, Interim Evaluation, p. A-7. The missing data issue is also used to support the use of the classroom as a unit-of-analysis by Marshall Smith in his Some Short Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation (Cambridge, MA: Huron Institute, 1973). For more extensive commentary on this reasoning, see Haney, Units-of-Analysis.

> One interpretation of this phenomenon is that the slowest pupils are most likely to be subject to attrition and/or pupils transferring into the program are academically behind the continuing (identic) pupils. However, since these post-test patterns were equivalent in FT and NFT classes, and since they were systematic across projects, a more likely interpretation is that identic pupils display a test-retest effect of approximately one item--on the average--over those pupils who migrate into the evaluation sample after pre-testing. The results, we feel, provide at least indirect evidence that use of identic pupil scores to estimate the classroom means does not produce subgroup bias (with the exception of the overall retest increment, which appears constant across all subgroups). This is particularly important, since control variable data exist only for the identic pupils. (p. A-10)

Since Emrick used the classroom as his unit-of-analysis, he was also concerned with the problems of attrition and missing data at the classroom level. He presents data showing the percentage of missing "aggregated classroom values on variable-by-variable bases" for each of the four cohort streams analyzed in the interim report. The proportion of classes with missing data ranged as high as 20% for some teacher variables and 16% for pupil outcome variables in one cohort stream (Cohort II EF). Emrick concluded that this

> was acceptable for the purposes of our interim analysis, [and] we proceeded to impute missing classroom-level values according to the following algorithm:
>
> --Data missing at the classroom level was imputed from the mean of the school at the appropriate subgroup classification (i.e., grade level by cohort by treatment condition).
>
> --If imputation within the school was unsuccessful, the next level of nesting--the district--was used to impute the missing values.
>
> --If imputation failed at the district level, then the overall subgroup mean was used to supply missing data values.

> It should be noted that only rarely was it necessary to proceed beyond the school level, and almost never beyond the district level, in the imputation procedure. As such, we feel the likely bias (most probably reduced variance and increase in Type I error rates) introduced by this procedure is small relative to its advantages of maintaining sufficient cases so that analyses could be performed. (p. A-11)

Despite their method of imputing scores to fill in for missing data, the authors of the SRI Interim Evaluation warned that the attrition problem boded ill for the future of the Follow Through evaluation.

> [P]reliminary evidence suggested that some rather serious problems will likely be encountered in the final stages of this longitudinal evaluation because of apparent patterns of differential FT/NFT attrition. Our current tracking data show that the half-life of the comparison samples is about two years. (That is, the size of the comparison sample reduced by one-half every two years, on the average). Thus, by the end of four years in this longitudinal study, only about one-fourth of the original comparison group is expected to be available for assessments. Although the attrition rate is considerably smaller for FT (half-life of about three years), it will be very difficult to draw valid inferences of effects, given what is likely to be a non-representative residual of an already biased comparison sample. (p. 320)

In these three major evaluation reports--Soar, Stallings 1972-73, and Emrick--the attrition problem did not figure very prominently. This was due mainly to two interrelated factors. First, all of these studies used the classroom as the basic unit-of-analysis. If some pupil post-test data were missing, this did not generally effect the analyses because it was assumed that other post-test data still provided an unbiased estimate of classroom average post-test scores.[3] The second factor helping to minimize the attrition problem in these studies was that each focused mainly on an analysis of data from a single school year. This allowed use of the classroom as a unit-of-analysis and also helped to minimize the attrition issue because student mobility is generally much less during a single school year than over more than one year.[4]

---

[3] It is unclear whether Emrick followed this assumption to the extreme case of allowing only one or two students' scores to calculate a classroom mean. As previously noted, Stallings required that "at least 20% of the children in the classroom had to have baseline test data," and Soar required at least seven pupil test scores per class.

The Attrition Issue in the First Phase of Abt Analyses. In subsequent evaluations, as Emrick et al. had predicted, the attrition issue achieved considerable prominence--though it was in rather unusual manner. The issue was prominent in the Abt I report mainly by its absence. In their evaluation of one-year kindergarten effects for cohort III, the Abt analysts provided scant detail on how their sample was created. They merely recounted:"The criterion for inclusion in the studies at the various levels of analysis is the availability of sufficient numbers of valid scores on children, their parents and their teachers" (p. III-2). At another point, Cline and colleagues state flatly that "our analytic subsets (we hesitate to call them samples) are not representative of any definable larger population."[5] In their parallel analyses at the school, class and pupil levels, the Abt investigators dispensed with the issue nearly as quickly:

> Class and child level studies were based on a subset of classes and children included in the analytic group of schools. A somewhat different set of inclusion criteria

---

[4] Exactly how Emrick used the classroom as a unit-of-analysis in calculating two-year effects is unclear. Apparently he employed the second year groupings of pupils (that is, first grade for the entering kindergarten streams, and second grade for the entering first streams), and used as pre-test covariates the average entering pre-test scores of those pupils (in K or first for the two streams respectively), even though they may not have been grouped together in the same classes during their first year in school.

[5] Cline, Abt IA, p. II-7. Parenthetical comment in the original.

> were applied to classes and children for those studies, so that the subsets produced at class and child levels are different from the school level groups both in numbers and in characteristics.[6]

In other words, the Abt analysts largely ignored the attrition issue in their first report by simply disclaiming that samples with post-test data were representative of samples with pre-test data, or any other samples.

In their second report the Abt analysts attacked the attrition issue more directly--at least apparently. For although the Abt II report devotes some 40 pages to the topic of sample development, those two score pages are easily the most murky among any of the many hundreds of hard-to-fathom pages in all the national evaluation reports. Essentially the Abt analysts seem to have compared various traced or longitudinal samples of cohort III-EK stream children with the total sample of cohort III-EK children who were tested as they entered kindergarten in the fall of 1971. The initially pre-tested C III-EK group was compared with two traced samples: one traced sample had test data at fall kindergarten and spring kindergarten (FK-SK), and another had fall kindergarten and spring first grade data (FK-S1). The Abt analysts compared these samples in a variety of ways, and from these comparisons drew several conclusions. First, with repsect to only the FK-S1 sample, they noted "it appears that there are differences among and within sponsors on the representativeness of the FT/NFT groups in the current [FK-S1] sample" (p. II-26). Moreover, "the current samples for several sponsors differ not only from the 1971 fall tested populations but also from last year's analytic samples [FK-SK] as well" (p. II-38). In conclusion, Cline and his colleagues noted:

---

[6] Cline, Abt IA, p. VII-53.

> The Sponsors' samples differ from each other in all possible ways. Sponsors vary in the extent to which the sample represents the population as well as the magnitude, direction, and type of FT/NFT mismatch. The Sponsors are working with different children from different families in different locations under different circumstances. Since the meaning of any Sponsor effect is inextricably tied to unique characteristics of the sample on which it is based, we cannot compare effects across Sponsors. (p. II-39)

The Abt analysts seemed to be saying exactly what they had said the previous year--that none of the samples they analyzed was representative of anything else--that within the same cohort III-EK stream, the FK-S1 sample was not representative of the initially pre-tested FK sample nor was it representative of the FK-SK sample analyzed in the Abt I report. Moreover, because of differential attrition in the Follow Through and Non-Follow Through groups for different sponsors, the Abt analysts reasoned, "we cannot compare effects across sponsors."

The conclusion was disturbing. The whole purpose behind the planned variation concept of Follow Through, to which the evaluation has been firmly latched since 1972, was to do what Cline and his colleagues recommended against--to compare effects across sponsors. Naturally, officals in the Office of Education were not pleased with this sort of conclusion. When the evaluation contract went out for bid in 1974, the number one question was whether sponsors have importnatly different effects. And before the contract was reawarded to Abt, the Office of Education wanted assurance on the comparability of the samples on which different evaluation analyses would be performed. In response to questions, the Abt evaluation team assured the Office of Education officials that during the second phase at Abt, a centralized evaluation team would make certain that any

> proposed analysis does not unnecessarily proliferate similar but different analytic subsets. By conducting as many analyses as possible on a single body of data selected in accordance with consistent criteria, we wish to enhance the cross-study comparability of this year's results and minimize the amount of pure data description that the reader must absorb.[7]

Feeling Better But Looking Worse. Despite Emrick's warning about the likely long-term severity of the attrition problem, and the conclusion of Cline and colleagues that attrition helped to make meaningful comparison of effects across sponsors impossible, a curious thing happened during the second phase of the evaluation at Abt. As the evaluation reached analysis of third grade outcome data, the attrition problem became worse, but no one seemed so concerned about its consequences. The natural question arises: Why did the problem look worse, but people seem to feel better about it? The answer seems to lie in the way the problem was examined.

In the Abt III report, the issue was viewed in two ways: 1) via a comparison of "pretest means for the FT and NFT groups prior to attrition (total sample) and after attrition (analytic sample)" and 2) by "a hierarchical regression analysis to determine the extent to which children who have not been retained (dropouts) are different in some important way(s) from those who are retained."[8] Since the hierarchical regression method was relied upon exclusively in the Abt IV report, let me describe only the first method here and postpone discussion of the second one until we get to the Abt IV report.

---

[7] N. Ames, R.B. Anderson, M. Cline, et al., "A Proposal to Analyze Selective Sets of Data Collected on the Follow Through Program. RPF HEW/OE 75-2. Responses to Technical Questions" (Cambridge, MA: Abt Associates, December 23, 1974). The final version of the Abt II report is dated May 1, 1975. However, a draft version of the Abt II report was distributed at the end of September 1974 and so was available prior to the negotiations for 1974 analysis contract.

[8] Abt III, p. A-48.

In their report III, the Abt analysts established several criteria for judging whether a particular Follow Through/Non-Follow Through comparison was trustworthy. One of them said simply that a comparison may be misleading "when the pre-test difference between the treatment and comparison groups being analyzed is greater than five points (about one-half standard deviation)" (p. A-71). The first method for examining the attrition issue in the Abt III report was simply to look at whether or not attrition pushed comparisons over the mistrustworthiness boundary of 5-points difference between pretests of Follow Through and Non-Follow Through groups which were to be compared. For the 51 sites in the cohort II-EK analytic sample, this comparison yielded the following findings:

> - In 39 sites FT and NFT differed by less than five points prior to attrition. In only five of these did the difference increase to more than five points after attrition.
>
> - In 12 sites, FT and NFT differed by more than five points at the outset. Attrition reduced three of these site differences to less than five points. (p. A-48)

As a result of this comparison and the hierarchical regression, the authors of the Abt III report concluded that "children in the analytic sample are representative of children who started in the Follow Through program at least in terms of pretest score and income."[9]

---

[9] Abt IIIA, p. A-49. The Abt analysts point out that when results are presented, "Sites which exceed a five point FT/NFT difference on pretest scores are 'greyed.' Attrition identified by the regression analysis is mentioned in the appropriate site discussion." The "grey-out" was used to indicate results that might not be trustworthy.

In the Abt IV analyses of cohort III-EK stream children, the attrition problem seemed to hit an all-time high: "approximately 70% of the FT and NFT children who were tested in the kindergarten year of cohort III-EK are not present in the analytic sample."[1] The figure was just about what Emrick predicted--a loss of about half of the originally tested sample every two years. When examined at the site level the attrition problem varied greatly--from a high loss of 84% of the originally tested Follow Through sample in the San Jose, California site, to a low of only 20% in the Montevideo, Minnesota site.[2]

The significance of the attrition problem in the Abt IV study was examined with "a hierarchical regression analysis to determine whether or not the FT/NFT within-site difference in rate of attrition was different across levels of pretest and income."[3] The analysis consisted of performing a multiple regression to see whether or not children's pre-test scores (or family income) were significantly related to their presence or absence in the analytical sample at the end of third grade. The regression relation of attrition on pre-test (or income) for particular sites was examined only if the whole set of on-site-pre-test (or site-income) variables across all sites proved significant.[4] On the basis of this two-stage test of attrition the authors of the Abt IV report concluded:

---

[1] Abt IVA, p. 82.

[2] Abt IVC, pp. 614, 812. In only seven of the 63 CIII-FK sites was attrition from FK to S3 less than 50 percent.

[3] Abt IVA, p. 83.

[4] The details of the hierarchical regression model for analyzing the attrition issue in the Abt IV study are given in an appendix, Abt IVA, p. 273. Note that despite the complexity of this approach it does not directly get at the issue of whether the predicted third-grade test scores would have been similar or different for the children who dropped out as for the children who stayed in the program and were tested at the end of third grade.

> In general, the examination of attrition leads to a conclusion that children in the Cohort III-K analytic sample are representative of children who started in the Follow Through program, at least in terms of pretest score and income. This suggests that the results of the pretest evaluation are generalizable to the Follow Through population to the extent that the original FT and NFT samples are representative of that population. (p. 84)

What happened? In any absolute sense the attrition problem seemed to be far worse in the Abt IV study than in earlier ones, yet the Abt analysts seemed far more sanguine about its consequences than either Emrick or Cline had been.

The answer, it seems, lies mainly with the sort of analysis used to test the significance of the issue. The hierarchical regression analysis employed exclusively in the Abt IV study to test for the significance of attrition is highly conservative because it employs a protected test of significance. The rationale for such a test is as follows. A series of tests of significance of attrition on a site-by-site basis would entail an accumulated error rate far in excess of the level of error assumed in any one test. Therefore, goes the protected test rationale, an overall test of the significance of attrition is needed to guard against finding significant differences which are in actuality due only to chance. What this strategy entails, however, is a tolerance of Type II errors (overlooking real differences) in an attempt to avoid Type I errors (finding differences as real when in fact they are due to chance).

The effect of the change in strategy seems straight out of Marshall McLuhan's the medium is the message, or in this case the analysis is the answer.[5] Yet the shift to the protected hierarchical test of attrition is not just a statistical sleight of hand. The bases for conducting a protected test of significance is well justified in the literature of data analysis.[6] The problem, it seems to me, lies in the disproportion between the conservatism of the tests applied to problems which would be convenient to do away with, and the liberality of tests applied to contrasts in which we are hoping to find significant "effects." For example, if a similar protected test strategy were applied in the search for Follow Through effects (first testing the overall significance of Follow Through versus Non-Follow Through differences, pooled across all sites and then only if they proved significant going on to test for significant FT/NFT differences for individual sponsors or sites), it seems likely that many if not most Follow Through "effects" would disappear.

There is, I think, no easy answer to this dilemma. I was a party to some of the discussion leading to adoption of the protected test strategy for analyzing the attrition issue and I must admit that at the time and discussed in isolation, the rationale seemed more persuasive to me. Nevertheless, I think it is important to acknowledge that the hierarchical regression strategy is somewhat conservative with regard to testing the significance of the attrition issue. Hence it is probably appropriate to be somewhat less sanguine on the attrition issue than this test alone leads us to be.

---

[5] If we use the alternative criteria employed in the Abt III report, for example, we get a somewhat different picture of the seriousness of the attrition problem in the cohort III-EK third grade data. In all, the Abt IV analyses of cohort III-EK data include 63 sites. In comparing original and traced (or analytic) samples of FT and NFT children, attrition pushes across the 5-point mistrusworthiness boundary ten sites in one direction (FT and NFT groups differ by less than 5 points in the original sample, but by more than 5 points in the traced sample) and two sites in the other direction. (See Abt IVC appendices.)

[6] For example, see Jacob Cohen and Patricia Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Hillsdale, NJ: Lawrence Erlbaum, 1975).

ANCOVA: Angel of Salvation or Temptress and Deluder?

> There is little question that most evaluation research in public service and social action areas will have to depend upon covariance analysis to approximate the controlled conditions of the laboratory experiment.
>
> - Edward Suchman[7]

A whole arsenal of techniques has been used to analyze the Follow Through evaluation data: regressed gain score analysis, correlation and partial correlation, cluster and factor analysis, multiple dimensional matching, discriminant analysis, multiple regression and path analysis. But among the multitude of methods, one clearly stands out--the analysis of covariance or ANCOVA. Emrick and his colleagues at SRI termed it their "principal statistical method for evaluating interim Follow Through impacts" (p. A-11) and Cline et al. called it their "fundamental mode of analysis" (Abt IA, p. V-5). Yet despite its widespread use in Follow Through and elsewhere, there still remains much doubt over whether ANCOVA is, in the words of a recent article on the technique, an angel of salvation or a temptress and deluder.[8] This section first reviews why such a controversial technique has been needed in the Follow Through evaluation, and then reviews the manner in which ANCOVA has been used in the evaluation. Three issues concerning the application of ANCOVA are treated: 1) the levels at which effects are estimated and for which type of "experimental subjects"; 2) the kinds of covariates which are selected to adjust for pretreatment differences; and 3) the manner in which the assumptions of ANCOVA are tested.

---

[7] E. A. Suchman, Evaluation Research, p. 105.

[8] L. J. Cronbach, D. R. Rogosa, R. E. Floden, and G. G. Prince, "Analysis of Covariance: Angel of Salvation or Temptress and Deluder?," Occasional Paper of the Stanford Evaluation Consortium, Department of Education, Stanford University, Stanford, CA. February 1976.

The Need for ANCOVA. If Follow Through had been a true treatment-control experiment with experimental subjects randomly assigned to treatment and control groups, analysis of outcome data might have been simple. Randomization would have helped guarantee the initial equivalence of the treatment and control groups so it would have been necessary only to compare the groups at post-test time in order to estimate the effects of the treatment. But there was no random assignment anywhere in the Follow Through program and FT and NFT groups have differed in numerous ways--as have the FT and NFT groups associated with different sponsors. As a result different groups of children in the program differ at post-test time not just because they received different "treatments," but also because they started out differently at pretest time. In the jargon of evaluation, treatment effects are confounded with initial differences.

The problem for the evaluation analyst then becomes how to sort out the differences among treatment groups at post-test time. After taking initial nonequivalence into account, the evaluator wants to know how much of the difference at post-test time is due to the treatment? To answer this question investigators at both Abt and SRI resorted to ANCOVA.[9]

---

[9] The problem of adjusting for initial differences was handled in other ways by Soar and Stallings. Soar calculated regressed gain scores for individual students first, then aggregated the individual regressed gain scores to get classroom scores and then used the classroom aggregates in subsequent analyses. In her analyses of test score data, Stallings used partial correlation and multiple regression to analyze classroom average scores.

ANCOVA has traditionally been viewed as a technique combining features of both analysis of variance and regression. More recently it has often been treated as a special case of a general linear model.[1] Formulations based on the two approaches differ slightly, but the essential features of ANCOVA and its limitations can be illustrated with the following model:

$$Y_{ij} = \overline{Y} + T_j + b(X_{ij} - \overline{X}) + e_{ij}$$

where

Y_{ij} = outcome variable Y for subject i in treatment group j;

$\overline{Y}$ = overall mean or average across all subjects on outcome variable Y;

T_j = treatment effect for treatment group j;

X_{ij} = value of the covariate or adjustment variable X for subject i in treatment group j;

$\overline{X}$ = overall mean across all subjects on covariate X;

b = regression coefficient showing the relationship between covariate X_{ij} and outcome measures Y_{ij} in the absence of any treatment;

e_{ij} = error associated with individual i in treatment group j.

---

[1] For example, G. Snedecor and W. Cochran, Statistical Methods, sixth edition (Ames, IA: Iowa State University Press, 1967) treat ANCOVA as a combination of analysis of variance and regression, whereas J. Cohen and P. Cohen in Applied Multiple Regression/Correlation Analysis for the Behavior Sciences (Hillsdale, NJ: Erlbaum, 1975) treat it as a particular type of the general linear model.

This model is considerably simplified over those upon which most Follow Through evaluation analyses actually have been based. For example, most analyses have used more than one covariate to adjust for the initial non-equivalence of treatment and control groups and sometimes several treatment effect terms have been included in the same model. Nevertheless, this simple model helps to illustrate a number of points about how ANCOVA has been used in the Follow Through evaluation.

The ANCOVA model assumes that a subject's score is the sum of an overall average, a treatment effect, some function of a covariate, and some amount of error. The model can be viewed as a special case of a more general linear model because the score Y_{ij} is a linear function of these elements. The model assumes that covariate values X_{ij} and errors e_{ij} are independent of each other and of the treatment. Also, it assumes a common relationship between the covariate and the outcome for all subjects in all treatment groups--which is often termed the assumption of homogeneity of regression.

The following pages recount how these aspects of ANCOVA have been handled in previous Follow Through evaluations. First we recap how previous analytical issues bear on the definition of treatment and experimental subject. Then we recount the way in which covariates have been selected and used in the Follow Through evaluation. The third section recounts how the assumptions underlying the ANCOVA model have been handled, and a final discussion sums up the way in which ANCOVA has been used in the Follow Through evaluation.

Experimental Subjects and Treatments. What is a Follow Through treatment and who or what is it that underwent any of the Follow Through treatments? The last part of this question, the identification of experimental subjects, is closely related to the units-of-analysis issue. As D. R. Cox has suggested, the unit-of-analysis should be "the smallest division of the experimental material such that any two units may receive different treatments in the actual experiment."[2] In the Follow Through evaluation, answers to the unit-of-treatment query have varied. Emrick maintained that "the locus of treatment in Follow Through is the classroom" (p. A-3). Cline and colleagues maintained that "The FT treatment is administered at the school level in the sense that a school is first selected for participation in the program, and then all of the eligible or desired classes are selected from within that school" (Abt IA, p. II-17).

The Emrick and Cline accounts differ, but have in common a quirk of language that reflects some of the ambiguity around the Follow Through "experiment." Though much of the language and many of the techniques of experimentation have been adopted by evaluators of Follow Through, references to experimental "subjects" are quite rare.[3] The rarity seems to stem from the quasi-experimental status of Follow Through. It is more than a non-experimental analysis of natural events, but not a true experiment where subjects are assigned to treatments. So the language used to describe Follow Through seems to go halfway, to refer frequently to experimental treatments, but only infrequently to experimental subjects.

---

[2] D.R. Cox, Planning of Experiments (New York: John Wiley & Sons, 1958), p. 2.

[3] The term "subject" crops up occasionally in Abt I and a few times in Elmore's dissertation, Follow Through: Decisionmaking, but is remarkably rare in the Follow Through evaluation literature.

Table 10: Levels of Effects Estimates and Subjects or Units-of-Analysis in the ANCOVA Used in the Follow Through Evaluation

Evaluation Report	Level at which Treatment Effects Estimated	Subjects or Units of Analysis
Emrick, Interim Evaluation, (pp. 74-86).	- Overall FT[1] - Sponsor - Project or Site	- Class
Cline, Abt I, (pp. VII-1-12, VII-34-35, VII-55-57).[2]	- Overall FT - Sponsor	- School - Class - Child
Cline, Abt II, (pp. V-1-5, VI-1-4).[2]	- Sponsor - Site	- Child
Stebbins, Abt III, (pp. A-51-75).[2]	- Site	- Child
Stebbins, Abt IVA, (pp. 65-69, 132).[2]	- Site - Sponsor[3]	- Child
Ferb, Abt IVD, Longitudinal Study, (pp. 7-10, 23-27).	- Sponsor preschool group[4]	- Child

---

[1] Although the point is not altogether clear, it appears that the Emrick study did not calculate overall Follow Through effects estimates directly, but instead indirectly as an aggregation of sponsor effects.

[2] Not all of the analyses in the Abt reports used the range levels or units reported here. The levels and units here refer to the main analyses in each of the Abt reports; specifically to the one-year kindergarten studies in Volume IA, to the kindergarten and two-year effects studies in Volume II, to the local and pooled comparisons in Volume III, and to the local and pooled comparisons in volume IV A.

[3] In the Abt IV A report, sponsor effects were not calculated directly, but rather as an aggregation of site-level effects.

[4] In the Abt IV Longitudinal Study effects were estimated for sponsor preschool groups, that is, for groups of children who previously had attended Head Start, had attended some other preschool, or who had not had any preschooling.

Yet whatever they've been called, the types of entities whose data have been subject to ANCOVA in the Follow Through evaluation have varied widely. This range, together with the levels at which Follow Through treatment effects have been estimated are shown in Table 10.

The trend shown here is ragged, but its general direction is clear. In terms of both level at which treatment effect is estimated and subject or unit-of-analysis, the Follow Through evaluation has been moving toward greater specificity. After starting out with the classroom as the unit-of-analysis, broadening to use three alternative units in the Abt I study, the evaluation has focused in recent years on the individual child as the appropriate "subject" or unit-of-analysis.

The level at which treatment effects are estimated has likewise been moving downward toward greater specificity, although the pattern here is somewhat more complex. The interim SRI evaluation estimated treatment effects at three levels: overall FT, sponsor, and project. In the first year at Abt, analysts dropped the site-level effects estimate and estimated effects at both the sponsor and overall FT levels. In the Abt II report, the overall FT effects estimate was dropped and the site-level effects estimate was added back. The report offered three reasons for returning to the site level for estimating treatment effects:

> The first reason is that there is a good deal of anecdotal evidence that very divergent socioeducational events are occurring across sites for a given Sponsor...

> The second reason for the site analyses is that the discrepancies between some of the kindergarten effects found with the traced two-year sample and those found for the one-year kindergarten sample appear in some cases to be related to site differences in the two samples...
>
> The third reason for the site analyses is that a substantial amount of heterogeneity of regression occurs in the overall Sponsor analyses. (p. VI-3)

This rationale once again illustrates the intertwining network of complications resulting from the deviations from the treatment-control experimental paradigm. The first reason meant simply that different things were happening at different sites even though they were associated with the same sponsor. In other words, the "treatment" appeared to be different at different sites. The second reason referred to the problem of attrition. Different sites appear to have different attrition rates and hence had different sized samples at post-test time. And the third reason--heterogeneity of regression--referred to one of the assumptions of ANCOVA, the assumption that the relationship between covariate and outcome measures is the same in treatment group and in the control group.

In the second phase at Abt, relatively little more was added to the discussions of treatment-level estimation and of unit or subject of analysis. The Abt III report recounted that "the selection of the level of analysis is based on our previous findings of sizable site variability and unique site circumstances, even within sponsor" (p. A-51). The same report repeats various practical considerations which forced use of the child as the unit-or subject-of-analysis.[4] The Abt IV report recounted mainly the same considerations on level of treatment estimation and subject of analysis,[5] but with one notable exception.

The exception came in the longitudinal study which was a special part of the Abt IV report. The longitudinal study attempted to investigate whether or not the effects of Follow Through might be different for children who had had different types of preschool experience. Specifically, it estimated effects separately for groups of children who had attended Head Start, had attended some other type of preschool or had not gone to any preschool at all. This fact presented problems, as the authors of the longitudinal study explained,

---

[4] Specifically, the Abt III report noted that the class could not be used as UA because "Classroom groups seldom stay intact from kindergarten through third grade" and the school could not be used since "few if any, sites have enough schools or classrooms to support a site-level analysis" (p. A-51). The Abt III report did, however, include a new discussion of the fact that since children in the same classrooms tend to resemble each other (i.e., children are not randomly assigned to classrooms), the use of the analysis in ANCOVA may lead to underestimation of the standard errors of the effects estimated with such analyses (see Abt IIIA, pp. A-51-53).

[5] Abt IVA, pp. 66-69.

> Clearly, the simultaneous exploration of the impact of Follow Through programs on children with different preschool experience in different sites would have been desirable. Such a study would clarify the stability of program impact across sites by preschool categories, would shed light on the differential impact of the programs for children with different preschool experience, and could add substantially to our understanding of implementation. Since programs are implemented within sites and since a program's impact on children is influenced by the success or failure of implementation, the site is a critical level of analysis. Furthermore, past Follow Through evaluations have demonstrated substantial variability in program effects among sites.
>
> However, the numbers of children within the site-by-preschool categories...are too small for the simultaneous exploration of both the preschool and site issue. Since other studies in the Follow Through evaluation...focus on site and since the effects of different preschool experience have not yet been adequately explored during the course of the prior evaluations, the latter issue is given precedence in this study and the site issue is ignored.
> (Abt IVD, pp. 8-9.)

The decision to use this analytical strategy, and the findings resulting from it, hold several implications for the Follow Through evaluation and the level at which we search for treatment effects. For now, however, let us note only that separating children into different preschool groups and calculating treatment effects for them separately is not the only way in which the preschool issue can be handled. In previous Follow Through evaluation analyses data on children's preschool experience had been used as a covariate in ANCOVA. So before returning to consider the level at which Follow Through treatment effects might best be calculated, let us stop to consider how and why different data have been included as covariates in the various Follow Through evaluation analyses.

Selection of Covariates. To adjust post-test scores so that treatment effects can be disentangled from initial nonequivalence, ANCOVA uses one or more covariates--the X described in the ANCOVA model above. The obvious sort of covariate to employ is some sort of pre-test. When only a pre-test is used as a covariate, there is no problem of covariate selection, and the technique can be considered as a sort of adjusted gain score analysis. But in Follow Through there was no simple pre-test for any of the outcome measures and so covariate selection becomes a problem.

The ANCOVA technique itself is largely indifferent to the sort of covariate employed. Indeed, it allows analysts great flexibility. To adjust for initial nonequivalence, they may use as covariates any variables on which treatment and control groups differ and which correlate--or covary--with the outcome measures. This flexibility has resulted in great diversity in the sorts of covariates employed in Follow Through evaluation studies. Emrick's evaluation of pupil test data employed the greatest diversity of covariates--apparently as many as 17 or 19 different ones although the exact ones used in different ANCOVA's is not entirely clear (pp. 76-86). What is clear is that the covariates used in the SRI analyses ranged widely: from the average number of pupils per classroom to the percentage of boys in each classroom, to the percentage of pupils in each classroom whose parents had skilled occupations.

The Abt I study employed a similarly diverse set of covariates. In the three parallel analyses using the pupil, class and school as units of analysis, Abt employed a total of 18 different variables as covariates (p. IV-6). However, only eleven of the variables are used as covariates in the pupil-level analysis, twelve in the analysis using the class as the UA and a different set of twelve covariates in the analysis based on the school as the UA.

The Abt II report is somewhat ambiguous on what covariates were used in which analyses, but it is apparent that a much smaller set of covariates was used than in the Abt I analyses. In their second year of investigation, the Abt analysts started with a list of around 50 potential covariates, but

> After careful examination of all the potential covariates only three survived for the overall analysis of sponsor FT/NFT differences: the WRAT as administered in the fall of the kindergarten year, family income, and mother's education.[6]

In the Abt III analyses, the number of covariates went back up. Three main analyses were performed for the Abt III study: 1) a local comparison analysis in which scores for Follow Through children in each site were compared with those for NFT comparison children in the same site; 2) a best match comparison in which FT children's scores in each site were compared with those for a best match comparison site (the comparisons were matched on the basis of ethnicity, average family income and average WRAT pretest score); 3) a pooled comparison in which FT children's score in each site were compared with all NFT children's scores pooled together into a single overall comparison group. In each comparison, ANCOVA was used to estimate treatment effects and in each at least ten covariates were used:

- Fall kindergarten WRAT;
- First language [English vs non-English];
- Family income [in dollars];
- Highest occupation in household [ten-point scale];
- Ethnic membership [two vectors defining White, Black or other];
- Sex [male or female];
- Entry age [in months];

---

[6] Abt IIA, p. IV-11. Appendix B to the Abt II report indicates that different of these three covariates were used in different ANCOVA's (Abt IIB, pp. B-22, B-24).

- Missing data code for WRAT;
- Missing data code for income; and
- Missing data code for occupation.[7]

In addition a set of interaction terms were entered as covariates in the pooled analysis for the Abt III study, "but due to limited degrees of freedom, were not entered in the local or best match analyses." These interaction covariates were:

- Fall kindergarten WRAT by ethnic membership (two vectors);
- Fall kindergarten WRAT by language;
- Sex by ethnic membership (two vectors);
- Sex by language;
- Entry age by ethnic membership (two vectors); and
- Entry age by language.[8]

Analyses for the Abt IV report differed in several ways from those for the Abt III study. The worth of the best match analysis was thrown into question when a second-best match analysis yielded results that were inconsistent with those of the best match analysis.[9] Also, the Abt analysts conducted several supplementary analyses to examine how specific changes in analytical methods would affect findings. Throughout these analyses, however, they retained use of the same covariates as used in the Abt III study as "core background variables." These were seven data variables (or eight if the two vectors for ethnicity are counted separately) and three missing data codes; for pretest, income and occupation. Also, as in the Abt III study, six interaction terms (or nine if each of the three terms involving ethnicity is counted as a separate vector) were included in the pooled analysis.

---

[7] Abt IIIA, p. A-57. An appendix table describing the exact ANCOVA model used for each of the three comparisons does not mention the missing data code for occupation, but its omission there was merely a typographical error. Personal communication with Linda Stebbins, May 1977.

[8] Abt IIIA, p. A-58.

[9] The Abt authors observed that the differences in results between the best-match and second-best match analyses "appeared to be unrelated to the relative degree of mismatch between the FT and the two NFT sites. This strongly suggests that the results are inconsistent, that they are dependent on the comparison group selected, and that the use of NFT groups with similar characteristics on matching variables produces different results." Abt IVA, pp. 251-253.

So the number of covariates used in the different outcome analyses has varied widely; from a high of around 18 in one of the analyses reported in the Emrick report to a low of only three in the Abt III study. In similar fashion, the space given to how covariates were chosen has also varied widely. Emrick et al. offered virtually no explanation for using the covariates they apparently did use, beyond a passing reference to "a series of preliminary screening analyses on potential covariables..." (p. A-15).

The Abt I report presents much more information on the ground for choosing covariates in a section titled Which Covariates? The Abt investigators took a comprehensive approach. They identified "six basic categories of extraneous factors [which may be]...confounded with FT and sponsor effects":

- pupil characteristics
- parent/family characteristics
- teacher characteristics
- class characteristics
- school characteristics [and]
- environmental characteristics.

The Abt analysts explain,

> These six categories of factors define a large universe of potential covariates for the analyses that follow. By no means were all of these variables measured or even measurable in FT. We selected our final list of 18 variables by applying five sequential constraints in approximately the following order: (1) the relevance of the variable as suggested by previous research, (2) our own thinking about the theoretical and methodological problems involved in the analyses, (3) the availability of the variable in the data base, (4) the requirement that a variable to be included correlate at a reasonable level with some outcome measure in the analyses, and (5) that the variable behave homogeneously across the various treatment groups; that is, that it not interact significantly with Sponsor and treatment variables.[1]

In their first report, the Abt analysts did not make clear how many potential covariates were considered before winnowing the list down to eighteen. In their second report, they became more explicit. Like the Abt I report, the Abt II study devoted a separate chapter to the issue of covariate selection. For their second study, the Abt analysts began the covariate search with a list of 44 variables of eight types.

- geographic and demographc variables (3)
- school district variables (8)
- school variables (6)
- characteristics of the FT/NFT children in the school (2)
- class variables (5)
- teacher variables (1)
- parent variables (8) and
- child variables (11).[2]

---

[1] Abt IA, p. IV-3.

[2] Abt IIA, pp. IV-2, IV-5. A list of potential covariates excluded indicates that some covariates were considered at different levels of aggregation.

Each potential covariate was then screened to determine whether it met all of the following criteria:

- lack of confounding with design
- presence of data
- correlation between outcome and covariate (at least .20)
- correlation between covariates
- sample size
- conceptual grounds
- nonsingularity of correlation matrix, and
- homogeneity of regression.

This screening process yielded the "final covariate set" for use in the Abt II report: the fall kindergarten WRAT, family income and mother's education, although only one or two of the final covariates were apparently used in some of the analyses.[3]

The Abt III report devoted relatively little space to the issue of covariate selection. It merely stated,

> The selection of covariates for each analysis was based on a number of considerations. First, we were limited to the variables on which we had data. Second, some variables did not correlate with dependent variables and consequently were not used. Finally, we were limited by the number of covariates allowable by the degrees of freedom available in the local, best match, and pooled analyses.[4]

---

[3] A fourth potential covariate, the fall PPVT score, also survived the Abt II screening. "However, it was used, in place of the Fall WRAT, only as a covariate for analyzing the spring kindergarten PPVT." Abt IIA, p. IV-11.

[4] Abt IIIA, p. A-57.

In the Abt IV report the space devoted to covariate selection went back up. The Abt IV approach to covariate selection was, nevertheless, very straightforward. After noting that "the main criterion for a good covariate is that it be correlated with the dependent variable," the Abt analysts simply examined the consequences of using different sets of covariates. First they tried including a set of dummy variables to control for variation between sites. They found that,

> [T]he amount of variance explained by the inclusion of a set of between-sites vectors *after* the inclusion of the demographic covariate set fluctuates from sponsor to sponsor and outcome to outcome, but often comprises about four percent of the outcome variance. This means that, in many cases, sites account for about ten or twenty percent of the predictable outcome variance. Similar conclusions hold for the pooled model. It should be noted that, on the average, sites (between-site vectors) account for somewhat more outcome variance than does treatment.[5]

Next, they tried using an alternative set of covariates which differed from the "core" covariate set in several important ways:

> The alternative set differs from the actual covariate set in that a variable indicating the ceiling value on income was added (allowing for a change in slope for these less disadvantaged children); ethnic membership and first language were combined into a four-vector variable; mother's education was added; the missing data codes for income, occupation, and education were combined into a single variable; and a term providing for the effects of a quadratic relationship between the pretest and each outcome was added.[6]

---

[5] Abt *IVA*, p. 85.

[6] Abt *IVA*, pp. 85-87. A table showing exactly how the core covariate set and the alternative set differ is given on page 86 of the Abt IVA report.

The Abt analysts then simply compared effects estimates resulting from use of the two sets. They found only three percent of the calculated effects "were changed enough to be reflected in a summary of effects," and even in those cases the changes in the magnitude of the treatment effects estimates were "very small."[7] As a result the analysts stuck to their use of the original variables as their core covariate set.

In another auxiliary analysis, the Abt analysts introduced preschool and preschool X pretest interaction terms as covariates and again considered how this changed patterns of effects estimates. Overall, inclusion of preschool and preschool X pretest covariates tended to make the results look somewhat better. For both pooled and local analyses, introduction of preschool and preschool X pretest covariate terms yielded more positive and fewer negative effects.[8] In an unusual set of linking analyses, the Abt analysts attempted to pinpoint the causes of the changes in effects estimates. Since explanation of these analyses gets quite complex very quickly, they will not be described here, save to note that as a result of these considerations, the Abt investigators rated the preschool adjusted analyses as primary and relegated the pooled and local analyses without preschool covariates to secondary status.

---

[7] Abt IVA, p. 87. Effects of these analyses were summarized in terms of whether they were positive, null, or negative. In order for an effect to be judged positive or negative, it had to be both statistically significant (at the .05 level) and at least one-quarter standard deviation of the outcome variable. See Abt IVA, p. 132.

[8] See Table 4-7, p. 100 and Table 4-10, p. 102, for summary tables of how the introduction of the preschool covariates changed results of the local and pooled analyses respectively.

So overall, the number of covariates used in the five outcome evaluations have varied widely; from around eighteen in the Emrick report, to around a dozen in most of the Abt I analyses, to three in the Abt II analyses, and back up to ten to eighteen in the Abt III and IV reports.[9] There has thus been relatively little stability in the way covariates have been chosen in the course of the Follow Through outcome evaluations. This seems to have stemmed from several factors. In the early years, the continuity problem seemed to derive partly from a sort of contract amnesia. In the rush of efforts to complete requirements for one-year contracts, there was little leeway to carefully consider all that has been done in preceeding years. In addition, the flux in covariates probably stems from different sets of "constraints" applied to potential covariates in different evaluation efforts. Five were listed in the Abt I volume and eight in Abt II. The marked drop in the number of covariates used in the second Abt report--from more than a dozen in Abt I to only three or four in Abt II--also is very likely related to the shift in levels of analysis. The Abt I report used the school and the class as units-of-analysis and outcome variables at these aggregate levels are likely correlated with potential covariates at the aggregate level more highly than they are with outcomes at the child-level used in the Abt II reports.

---

[9] Ten covariates (seven data variables and three missing data codes) were used in the local analyses in the Abt III and IV studies, sixteen in the Abt III and IV pooled analyses (including six interaction terms) and eighteen in the Abt IV preschool adjusted pooled analysis.

In the Abt IV study, analysts took the more straightforward approach of trying out different sets of covariates to see how this changed results--both in terms of proportion of variance explained and in terms of resulting effects estimates. When effects estimates differed from the different analyses, they attempted to assess which ones were more trustworthy. For example, they judged that analyses adjusting for differing levels of preschooling were better than those that did not.

This particular finding presented a bit of a predicament, however. As a result of their analyses for volume IV, the Abt analysts judged that the analyses using preschool covariates (that is, the preschool adjusted local analysis and the preschool adjusted pooled analysis) were more trustworthy than the analyses which did not include preschool covariates (the simple local and pooled analyses). Yet the Abt III report had been based entirely upon analyses which did not adjust for preschool. So one major finding of the Abt IV report--though nowhere stated outright--was that all three main analyses for the Abt III report might not be trustworthy.[1] Because of this, OE awarded a continuation contract to Abt in 1977, to go back and conduct preschool adjustments on the three cohorts streams for which effects estimates had previously been calculated without adjusting for preschool attendance.[2]

---

[1] Recall that the three main analyses in the Abt III study were the local, pooled, and matching analyses. The matching analysis, as previously noted, was abandoned in the Abt IV study after a second-best match cast doubt on whether it could be trusted.

Although the Abt III analyses did not take preschool into account, the reporting of the Abt III results did so a posteriori, by "greying-out" results for sites in which Follow Through and comparison groups differed in terms of proportion of children attending preschool by 50% or more.

Contrasting the Abt IV strategy for picking covariates with the approach taken in the Abt II report raises one other question of interest. In the Abt II study, family income and fall WRAT were excluded from some analyses of first grade outcome scores because they lacked homogeneity of regression with outcomes.[3] This sort of consideration was not even mentioned in the Abt IV study as a grounds for selecting covariates. The question which naturally arises then is how this and other assumptions of ANCOVA were tested in the different studies.

ANCOVA assumptions. The assumptions underlying the ANCOVA model can be posed in a variety of forms. The assumptions receiving most attention in the Follow Through evaluation concern the homogeneity of regression relations, the reliability of covariates and the independence of error or residual terms.

There is no well-developed etiquette on how to apply a statistical model like ANCOVA and hence no easy rules for how to test the assumptions underlying the model. As a result there has been wide variety in the ways

---

[2] In the portions of the Abt IV report completed in April 1977, preschool adjustments had been included only in analyses of cohort III-EK data. The 1977 continuation contract allowed Abt analysts to recalculate effects estimates for the cohort II-EK and EF streams and the cohort III-EF stream using preschool adjustments. These new preschool adjusted analyses are reported in Abt's IVE and IVF volumes--which are not covered in the technical history.

[3] Abt IIA, pp. IV-10-11.

ANCOVA assumptions have been treated over the course of the national evaluation. There seems to have been widest variety in the handling of assumptions concerning independence of error and treatment terms in the ANCOVA model. In technical appendices both the Emrick report (p. A-13) and the Abt I study (Abt IB, pp. IV-1-20) discussed the assumption that error terms are independently distributed, but neither provided information regarding empirical tests of the assumption. The Abt II report reiterated the Abt I discussion on this point (Abt IA, pp. B-1-22), but again provided no indication as to how or if the independence of errors assumption had been empirically tested.

The Abt III report couched discussion of independence assumptions in somewhat different terms, saying that covariates should be "uninfluenced by treatments."[4] Abt analysts, then argued that since demographic variables are not susceptible to direct influence by treatment, this assumption need be tested only with regard to the WRAT pretest score. They continue:

---

[4] Note that the assumption that covariates are uninfluenced by treatments is not quite the same as assuming that covariate values are independent of treatments. Historically ANCOVA was first used not to adjust for initial pretreatment differences but to increase precision in randomized experiments (G. Snedecor and W. Cochran, Statistical Methods, p. 419). In such cases it would be appropriate to assume that covariates are uncorrelated with treatments. However, in using ANCOVA to adjust for initial differences in treatment groups, covariates are obviously not uncorrelated with treatments. Therefore, the weaker assumption, that covariates are uninfluenced by treatments, is appropriate when ANCOVA is used to adjust for bias in nonrandomized situations.

> The WRAT was administered by SRI early in the kindergarten year, usually within the first six weeks of school. Although the WRAT is sensitive to early learning, it does not seem likely that the instruction received during these first weeks (which are generally spent "getting settled") would substantially affect WRAT scores. But if the WRAT is influenced by the first few weeks of treatment, one might expect pretest adjustments to handicap the FT children. To test this, we removed the WRAT from the covariate set and reran the local analyses. The results of these "no-WRAT covariate" analyses do not differ in any important ways from analyses which included WRAT as a covariate. We conclude from this comparison that the WRAT is probably not hindering our analyses of program effects.[5]

The Abt IV report repeated the first two sentences of this argument almost exactly word for word. But then there was a switch,

> On the other hand we have no evidence to suggest that early learning which could be measured by the WRAT did not take place for some children. The reader especially concerned that such a phenomenon did occur may wish to pay special attention to our analyses that omit the WRAT.[6]

The change in tone was subtle, but clear: from "reader believe," to "reader beware." When the wary reader turned to the analysis omitting the WRAT from the preschool adjusted local analysis, the overall pattern of effects estimates remained very closely the same: out of 930 estimated effects 7 percent were positive, 80 percent were null, and 13 percent were negative both before and after the pretest covariate was omitted. The catch, however, was that they were not the same ones. When the pretest covariate was dropped from the analysis, around one-third of the positive effects estimates became null and an equal number went from null to positive. Similar reversals happened with negative effects estimates. Overall, there was an 88 percent agreement from the two analyses as to positive, null or negative effects estimates.[7]

---

[5] Abt IIIA, pp. A-58-59.

[6] Abt IVA, p. 98.

The homogeneity-of-regression-relations assumption has undoubtedly received the widest attention in the Follow Through evaluations of any of the assumptions underlying the ANCOVA model. This assumption says simply that the same regression relationship between the outcome variables Y_{ij} and the covariates X_{ij}, expressed as b in the model, holds among all the treatment and control groups in the ANCOVA.[8] Emrick mentioned this assumption but recounts that the "within-cell frequencies were frequently too small to enable a test of this assumption" (p. A-15). In the first two Abt reports, the homogeneity-of-regression assumption was given much more prominence. In both studies it was one of the constraints used to screen lists of potential covariates. In the Abt I study, analysts apparently tested selected covariates for regression homogeneity by entering interaction terms into their analyses.[9] This point is a bit unclear, however, because the Abt I analysts also report: "Our samples are too small to permit us to introduce appropriate interaction terms in the model to adjust for covariates heterogeneity" (Abt IA, p. IV-3). How the assumption was handled becomes even murkier in the Abt II report. Although one of the criteria for covariate selection was supposedly homogeneity of regression, a number of results are not reported because of "heterogeneity of regression" (e.g., p. VI-118).

---

[7] Abt IVA, p. 89. Recall that the criteria determining that an effect be non-null were that it is both statistically significant at the .05 level and at least one-quarter standard deviation of the outcome variable. The 930 effects estimates were for 62 cohort III-EK sites and 15 "outcome" variables. Effects were estimated separately for eleven scores derived from the MAT (for the seven MAT subtests, plus Total Reading, Total Math, Language A and Language B scores), for the Ravens, the Coopersmith, and IARS positive, and the IARS negative.

[8] Cronbach et al. suggest in "Analysis of Covariance: Angel of Salvation or Temptress and Deluder?" that the assumption of homogeneity of regression relations is not absolutely necessary to use ANCOVA. However, their formulation simply transforms the problem into a more general one of model specification--and as the wide range of covariate sets used in Follow Through evaluations shows, the model underlying the analyses of Follow Through outcomes is not at all well-specified.

[9] Abt IB, pp. IV-18-20; Abt IA, p. V-3.

In the Abt III study the handling of the homogeneity assumption was considerably clearer--but the problem was still tinged with grey.

> Homogeneity of regression was tested in each local comparison by examining separately each covariate and its potential interaction with treatment. We thus addressed the question of whether the FT/NFT within-site difference on a given post-test measure differed across levels of a given covariate. For each sponsor, each covariate, and each posttest measure, a regression analysis was performed which predicted a given posttest as a function of a given covariate, a set of orthogonal between-sites contrasts, a set of orthogonal within-site contrasts, the interactions of the covariate with the between-site contrasts, and the interactions of the covariate with the within-sites contrasts. We then determined whether the last set of interaction terms (covariate by treatment within-site) explained a significant amount of variance in the dependent variable ($p<.01$). These individual terms have been identified and...problems of this type contributed to our decision to grey-out certain effects.[1]

So some of the Abt III effects estimates were "greyed-out" to indicate that they were mistrustworthy on account of homogeneity of regression.

In their Abt IV study, the analysts began with the same test for homogeneity of regression as had been reported in the Abt III study. They found:

---

[1] Abt IIIA, pp. 59-60.

> There were some sites in which the FT and NFT regression lines of some outcome on a given covariate were nonparallel when assessed via this method, a result which may have occurred because the technique is not wholly satisfactory since it only assesses the effects of one covariate at a time. A second method was therefore developed with the advantage of examining all covariates at once; however, it is restricted in that it only contrasts a given site with all others of the same sponsor .... In the second method, we partially obscure between-site variance in order to gain a clearer understanding of the FT/NFT outcome-covariate relationship when all covariates are operating. Selective application of this technique to those covariate by treatment within-site interactions found to be significant earlier did not confirm the results of the first method and, in fact, showed that in no cases..were FT and NFT regression surfaces nonparallel. We interpret this as enabling us to proceed with the Local Analysis, unencumbered by the problems of heterogeneous regression for treatment and control groups.[2]

Hmmm. Once more a potential problem seems to fall in face of a two-stage test of its significance. So it seems the analysis is again the answer. There is no Hoyle on homogeneity testing, and this approach clearly follows the general strategy set out by several authorities in this field.[3] But once again the test to which the heterogeneity problem has been put seems far stiffer than the tests applied in the search for treatment effects.[4]

---

[2] Abt IVA, p. 97-98.

[3] For example, Fred N. Kerlinger and Elazar J. Pedhazur, Multiple Regression in Behavioral Research (New York: Holt, Rinehart & Winston, 1973), pp. 276-71, or Jacob Cohen and Patricia Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Hillsdale, NJ: Lawrence Erlbaum, 1975), pp. 343-45.

[4] The Abt IV test for heterogeneity-of-regression relations seems particularly stiff in two respects. First, it is a two-stage test in which a covariate must pass (or fail, depending on one's perspective) both tests in order to be declared heterogeneous. Second, the tests were applied in a manner analagous to the local analysis, but not in one similar to the pooled analysis (compare Abt IVA, pp. 96-98 with pp. 256-57 and with pp. 258-59).

Complete reliability of covariates is assumed with ANCOVA because a large degree of error in covariates can prevent adequate adjustment for initial nonequivalence between treatment and control groups. This assumption was relatively abstruse and little recognized until 1970, when Donald Campbell and Albert Erlebacher published an influential article criticizing the Westinghouse-Ohio evaluation of Project Head Start. The article was entitled "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful."[5] Campbell and Erlebacher argued that in evaluations of programs like Head Start and Follow Through, where "treatment groups are selected because they are disadvantaged, comparison groups tend to be relatively better off." Hence when fallibility in covariates causes underadjustment for initial nonequivalence, the estimated treatment effects will appear smaller than warranted or perhaps even negative They concluded, "It is tragic that the social experiment evaluation most cited by presidents, most influential in governmental decision making, should have contained such a misleading bias."[6]

Given this background, it is not surprising that the reliability of covariates issue has been receiving increasing attention in the Follow Through evaluation. In the Emrick report, it was noted that ANCOVA requires covariates that are measured without error. Emrick et al. argued, however, that

---

[5] Donald T. Campbell and Albert Erlebacher, "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful." In J. Hellmuch (ed.), Compensatory Education: A National Debate. Vol. 3. New York: Bruner Mazel, 1970. pp. 185-210. The Campbell-Erlebacher article was reprinted in E. L. Streuning and M. Guttentag (eds.), Handbook of Evaluation Research (Beverly Hills, CA: 1975) pp. 597-617.

[6] Campbell and Erlebacher, p. 203. Campbell and Erlebacher go on to call for true randomized social experiments.

> [S]ince our data are aggregated to classroom-level variables, reliability estimates of all covariables are such that this correction procedure would have virtually no net effect. All covariables for pupil and parent analyses have estimated reliabilities (where estimable) in excess of .95, and currently, the reliability of teacher covariables is unestimable and thus must be assumed to be error-free.[7]

In the Abt I report, Cline and colleagues identified the issue as "fallibility of measures."[8] They attempted to do something about the problem by adjusting one of their ANCOVA's for covariate unreliability. In the child level analysis, they apparently attempted one analysis in which the Follow Through effects estimates were calculated after adjusting for "initial inequalities, as well as the unreliability of the most important and fallible covariable, initial achievement level as measured by the fall WRAT" (p. VII-55). Unfortunately, the Abt analysts did not describe exactly how they accomplished this adjustment nor did they discuss the consequences of this effort in comparison to other effects estimates. Inspection of their results indicates that the adjustment for unreliability of covariates did not make much difference on effects estimates--in either magnitude or direction.[9]

The problem of covariate fallibility was not treated, nor even mentioned in the Abt II report--apparently another casualty of contract amnesia. The issue revived again in the Abt III report, but was dispensed with in somewhat the same fashion as it had been handled in the Emrick report. The Abt analysts wrote that the requirement of reliable covariates

---

[7] Emrick, Interim Report, p. A-34.

[8] Abt IA, p. V-3.

[9] The only exception was an effects estimate on the WRAT outcome for the Bank Street model in which case the adjustment apparently made a difference of half a standard deviation and reversed the direction of the estimate (Abt IA, p. VII-78).

> is easily met for most of the covariates in this study. Variables such as sex, ethnicity, income, occupation, education, language, and age are all measured with a minimum of error. It is only the pretest which poses a potential problem. The reliability of the pretest was calculated on each Follow Through sponsor-level sample by a measure of internal consistency (coefficient alpha) and is on the order of .90 across these samples.[1]

In the Abt IV report the covariate reliability issue was given fuller attention than anywhere else in the Follow Through evaluation. The Abt analysts began with a reiteration of what they had written in the Abt III report; namely that all of the covariates besides WRAT pretest are "presumably measured with minimal error," and that the coefficient alpha reliability of the pretest "is on the order of .90."[2] However, instead of doing away with the issue on this basis, the Abt investigators went on to consider how their analyses might be corrected for unreliability of WRAT scores. They first noted that even in the case of ANCOVA with a single covariate, there is disagreement in the literature as to which kind of reliability coefficient (internal consistency or test-retest) should be used to adjust for covariate unreliability. The situation is even more complex in the ANCOVA with more than one covariate, noted the Abt authors, because "the solution to the problem in the multiple covariate case (even if only one of the covariates is unreliable) is not clear."[3] In view of this uncertainty, they decided simply to try out a method proposed by Jacob and Patricia Cohen.[4] The Abt analysts recount:

---

[1] Abt IIIA, p. A-60.

[2] Abt IVA, p. 93.

[3] Abt IVA, p. 94

> Applying their method to the present case, we desire to correct only for the effects of unreliability in the pretest. The procedure involves correcting the correlations of the unreliable covariate(s) with each other covariate and the outcome for attenuation due to unreliability by dividing each correlation by the square root of the estimated reliability of the covariate. In addition, the covariate standard deviation is corrected by multiplying the observed standard deviation by the square root of the covariate reliability. ...Given the disagreement in the literature as to the most appropriate reliability estimate and the fact that we do not have a direct measure of the pre-post correlation for the fall kindergarten WRAT, the Local Analysis was run for Cohort III-K (on the subsample of pupils having preschool information) using three separate values spanning the range of potential estimates. The values selected were .6, .8, and 1.0, the latter being the equivalent of not correcting the pretest. Our interest is in determining whether the analytic results obtained by correcting for pretest unreliability (.6 or .8) differ from results obtained from an uncorrected analysis (1.0). Summaries of the results of the three analyses... indicate that lower reliability estimates lead to movement of effects from the null category. Correction for unreliability tends, furthermore, to make the results less favorable to Follow Through: without correction, 791 (85 percent) of the results are either positive or null; this number drops to 696 (75 percent) when corrected for a .80 reliability estimate and to 647 (70 percent) when corrected for a .60 reliability estimate.[5]

---

[4] Jacob Cohen and Patricia Cohen, Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (Hillsdale, NJ: Lawrence Erlbaum, 1975). The Cohens acknowledge, and the Abt analysts recount, that their method of adjusting for covariate unreliability has not been mathematically proved and "rests on no more than the judgment of the present authors and some of our colleagues" (Cohen and Cohen, p. 373).

[5] Abt IVA, p. 94. The exact consequences of using the adjustments are given in tables 4-5 and 4-6, and appendix table A4-2 of the Abt IV report.

If unreliability of covariates does tend to mistakenly make Follow Through compensatory education programs look harmful, it is clear that this effort to rectify the situation did not prove out. More such analyses doubtlessly could have been performed,[6] but it is doubtless too that they would not have yielded any firmer answers to the question of what Follow Through program effects have really been. So in closing, let us pull back from these specific issues and consider more broadly how ANCOVA has been used in the national Follow Through evaluation.

---

[6] For example, it is possible that this attempt to adjust for covariate unreliability tended to make overall results look worse because it was performed with an analysis that did not take preschool into account. On this basis one might argue that adjusting for WRAT unreliability should be done with the preschool adjusted analyses. Also, the assumption that demographc covariates (particularly income, occupation, and education) are measured with a minimum of error, could be challenged. Hence it could be argued that analyses adjusting for possible unreliability in these covariates should be performed as well as those adjusting for unreliability in the WRAT pretest.

* * * * *

ANCOVA: Summing Up. Is the ANCOVA a savior or a temptress and deluder? The query posed at the start of this section has been receiving considerable attention lately. Evidence is mixed but informed sentiment seems to be shifting to the view that in quasi-experiments like Follow Through, where treatment and control groups do not start out as essentially the same, the ANCOVA may often be more a deluder than a savior. For example, in an investigation of a variety of linear adjustment techniques including ANCOVA, Anthony Bryk and Herbert Weisberg recently concluded that "traditional adjustment strategies cannot be trusted to adjust adequately for nonequivalent growth systems across treatment and control groups."[7]

In the context of Follow Through the question of whether ANCOVA is a savior or deluder is still open, but it seems to have narrowed some. ANCOVA has not proved to be an easy savior, but neither has it yet been shown to be an outright deluder. If anything at all, it seems to have become a sort of grand diversion and great complicator.

The ANCOVA technique has been used in many ways in the Follow Through evaluation--with four different levels of treatment estimation, and three kinds of entities as "subjects," and literally dozens of kinds of covariates. The Abt I report even wandered off into a 50-page treatise on ANCOVA and the general linear model.

---

[7] Anthony S. Bryk and Herbert I. Weisberg, "Use of Nonequivalent Control Group Design When Subjects Are Growing," Psychological Bulletin, forthcoming, 1977. From a very different sort of investigation Cronbach, Rogosa, Floden and Price reached a similar conclusion regarding ANCOVA. "The sufficient conditions for an accurate description of the treatment effect [using ANCOVA in a quasi-experimental or 'two population experiment'] will not be met in practice" (p. 33).

The experimentation with the application of the ANCOVA technique has contributed to a number of discontinuities in overall analyses. For example, the Abt II study added a site-level treatment estimation to the sponsor level of the Abt I study, but the chapter of the Abt II report dealing with sample attrition analyzed the problem at only the sponsor level.[8] In similar fashion, the Abt III and Abt IV analyses estimated treatment effects at the site level, but dispensed with problems of attrition and heterogeneity of regression relations mainly at the overall level.

The varied application of ANCOVA to the Follow Through evaluation data has also resulted in a variety and complexity of analyses which tends to leave the end results quite impenetrable. In the Abt I report, the ANCOVA was employed to estimate effects at two levels (sponsor and overall) with three alternative units of analysis (school, class and child) for eight different outcome measures (WRAT, MAT, listening, reading and numbers, gumpgookies, locus of control positive and negative, and absence). This resulted in such a bewildering array of effects estimates, often contradictory at the three levels of analysis, as to make synthesis virtually impossible. As the Abt investigators commented, "there are too many studies to synthesize into a single set of effects patterns."[9]

---

[8] For example, see Abt IIA, pp. II-24-II-39. The sample development chapter described numbers of children in the various samples on a site-by-site basis, but examined the problem of attrition bias in terms of pretest scores and SES measures solely at the sponsor level.

[9] Abt IA, p. IX-1.

One of the reasons for the complexity of the Follow Through evaluation analyses is of course that the world can be immensely complex. Consider, for example, just the dimensions which the Abt IV report, at one point or another shows to be determinants, or at least correlates, of variations in effects. First of all, the study includes 17 sponsors, for some of whom effects seem to vary. Then there are around 100 different sites included in the Abt IV study. Effects seem to vary also according to cohort-stream (cohorts II and II entering first and entering kindergarten streams); according to three types of children's preschooling (Head Start, other preschool, no preschool); to community size (Philadelphia-New York, other big cities, medium and small cities, and rural communities); to period over which effects are estimated (early or FK to S1; late or S1 to S3; and overall or FK-S3); to outcome measure (fifteen different outcome measures: the Ravens, the Coopersmith, the IARS positive and negative and the eleven scores derived from the MAT); and according to site income level (high or low). Now if effects could be estimated independently along all of these dimensions of variation, we could calculate around 7,300,000 different estimates.

In actuality, of course, effects cannot be calculated independently along all these dimensions. Effects cannot, for example, be estimated separately for sites and community sizes because sites are fixed in particular communities. In the jargon of experimental design, sites are confounded with community size. Since many of the other dimensions mentioned above are also confounded with each other, the actual number of effects which can be calculated from the Follow Through data is fortunately somewhat smaller than 7 million.

Most of the Abt IV analyses focused on just one cohort-stream (CIII-EK) and only 62 sites. Since effects were calculated separately for fifteen outcome measures, this meant that each of the Abt IV site-level analyses yielded 930 effects estimates. The Abt IV longitudinal analysis also showed that effects estimates varied by period over which they were estimated, and by preschool group. Considering only these four dimensions which can be considered independently, we have a possibility of 62 X 15 X 3 X 3 = 8370 different possible effects estimates. However, as the authors of the longitudinal study noted, "the numbers of children within the site-by-preschool...categories...are too small for the simultaneous exploration of both the preschool and site issue."[1]

So the Abt analysts ended up by alternatively looking at the Follow Through evaluation data from different angles, cutting through the data in different ways. One time it was disaggregated by groups of children with different kinds of preschool experience but with children in different sites treated together; another time it was examined at the site level, but with preschool groups lumped together. "Effects" estimates can be calculated from each of these perspectives, but the question remains, which of them are real, and which are mere shadows of real effects which only show up from other perspectives?

The answer is that there simply is no way to be sure. The ANCOVA's used in the Follow Through evaluation are what John Tukey would call a form of exploratory data analysis rather than confirmatory data analysis.[2]

---

[1] Abt IVD, p. 9.

[2] John W. Tukey, Exploratory Data Analysis, Limited preliminary edition (Reading, MA: Addison-Wesley, 1970), Preface.

The analyses of the Follow Through data do not represent precise statistical analyses of the sort suggested by the ANCOVA model described at the start of this section. The Follow Through data simply do not fit terribly well with the assumptions of the formal statistical models on which analyses have been based. Tests of these assumptions have I think, sometimes evolved in such a way as to show that the contrary is the case, but in data analysis of this sort, the point may be somewhat irrelevant. This view has been aptly expressed by Jacob and Patricia Cohen.

> The behavioral scientist relies very heavily on the fruits of the labors of theoretical statisticians. They provide guides for teasing out meaning from data, limits on inference, discipline in speculation. Unfortunately, in the textbooks addressed to behavioral scientists, statistical methods have often been presented more as harsh straightjackets or Procrustean beds than as benign reference frameworks. Typically, a method is presented with some emphasis on its formal assumptions. Readers are advised that the failure of a set of data to meet these assumptions renders the method invalid. All too often, the discussion ends at this point. Presumably, the offending data are to be thrown away.
>
> Now this is, of course, a perfectly ridiculous idea from the point of view of the working scientist. His task is to contrive situations that yield information about substantive scientific issues--*he must and will analyze his data*. ...Data analysis accepts "inadequate" data, and is thus prepared to settle for "indications" rather than "conclusions." It risks a greater frequency of errors in the interest of a greater frequency of occasions when the right answer is "*suggested*." It compensates for cutting some statistical corners by using scientific as well as mathematical judgment, and by relying upon self-consistency and repetition of results. Data analysis operates like a detective searching for clues rather than like a bookkeeper seeking to prove out a balance.[3]

---

[3] Cohen and Cohen, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, p. 15.

From this perspective, it becomes easier to understand why different investigators have fit the bits of evidence deriving from the various ANCOVA's and other analyses of Follow Through data together in different ways. Different detectives may fit clues together in varied ways, to yield conclusions which may or may not differ in particular respects.

Consider, for example, how the pieces of evidence regarding Follow Through effects have been handled in the second phase of analyses at Abt. Recall that in the Abt III and IV studies, effects estimates were calculated at the site level (and in the Abt IV longitudinal study at the sponsor-preschool group level). The problem then was how to aggregate effects estimates back up to the sponsor level which was, after all, the level at which Follow Through treatments were nominally defined. In the Abt III study, effects estimates were simply not aggregated at all. As recounted in chapter I, a draft of the Abt III report contained a summary chapter 19, which reported sponsor aggregated effects. However, when investigators became concerned over the preschool issue, the summary chapter was dropped from the final version of the Abt III report to prevent premature conclusions from being drawn from the report. In the detective metaphor, the investigators were worried that an important clue to conclusions regarding Follow Through effects had been omitted and so they simply declined to draw overall conclusions.

In the summary to the Abt IVA volume, effects were summarized for each sponsor as the proportion of net positive effects among all trusted effects.[4]

---

[4] Abt IVA, p. 133.

In the Abt IV longitudinal study, summary of trusted effects to the sponsor level was accomplished in two ways: by calculating the proportion of positive effects estimates for each sponsor; and by calculating the proportion of non-negligible effects for each sponsor.[5]

These varieties of means for summarizing Follow Through results, together with the varied analyses on which they are based, show that it is delusion to believe overmuch in the results of any one of the Follow Through evaluation analyses, or even in the summary of any one set of analyses. Any one ANCOVA on the Follow Through evaluation data likely would be a deluder if viewed in isolation. Whatever saving power there is in this technique, it is only to be derived from a balanced consideration of results from the varied perspectives from which the Follow Through evaluation data can be viewed.

---

[5] Abt IVD, pp. 73-74. The difference in the Abt IVA and Abt IVD methods of aggregating effects is, of course, that the IVD methods tend to make results look somewhat more favorable than the IVA method which entails subtracting negative effects from positive ones.

IV. ANALYSES OF PARENT INTERVIEW AND TEACHER QUESTIONNAIRE DATA

> [E]fforts were being made to broaden the scope of the assessment...[but] the purpose for each of these additional efforts was unclear and some difficult conceptual analysis and development had to be undertaken. For example, obviously one ought to be interviewing parents as well as teachers; but what did one want to learn from these interviews? Everyone could agree that the attitudes of parents and teachers had some influence on children and on classroom outcomes but it was much more difficult to agree on a model or theory about these influence processes or on a statement of hypotheses that were to be tested with the aid of these data.
>
> -Gilbert Krulee, describing
> evaluation planning in 1968-69.[1]

The bulk of analyses in the national Follow Through evaluation have focused on outcomes for participating children as measured by the several sorts of test data. These analyses were reviewed in the last chapter. Another area of analysis has been that of parent interview and teacher questionnaire data. The analytical methods employed in this realm are largely different than those employed with the pupil test data and so are reviewed separately in this chapter.

The chapter has three sections. The first two describe the parent and the teacher data respectively. Each section provides a short review of the history of the data-gathering instruments prior to reviewing the analyses of the parent and teacher data. The third section briefly recounts uses of the parent and teacher data outside major evaluation studies and sums up the fate of the parent and teacher data in the national evaluation. The instrument histories and the reviews of analyses of the parent and teacher data reveal that the problems Krulee identified in 1973, dating back to the very first year of the national evaluation, have never been satisfactorily resolved.

---

[1] G. Krulee, Organizational Analysis (April, 1973), p. 207.

Evaluation of Parent Interview Data

> EVERY FOLLOW THROUGH PROJECT MUST PROVIDE FOR SIGNIFICANT PARENT PARTICIPATION IN ALL ASPECTS OF THE PROJECT.
>
> -Draft Follow Through Program Manual, February 24, 1969.

Follow Through was originally authorized under the community action title of the Economic Opportunity Act of 1964, the title which required programs to be "developed, conducted and administered with the maximum feasible participation of groups referred to...."[2] As part of the community action legislation, Follow Through, like Head Start, emphasized parent involvement. Indeed, the Follow Through program manual decreed that "at least four major kinds of parent participation are necessary for an effective project:

- Participation in the process of making decisions about the nature and operation of the project through frequent meetings of a Policy Advis ory Committee [PAC] and other parent groups;
- Participation in the classroom and school as paid employees, volunteers, or observers;
- Provision for regular home contact by Follow Through staff;
- Parent educational and community activities which parents have helped develop.[3]

When proposed rules and regulations for Follow Through were published in 1974, parental involvement again drew considerable attention. Among fifteen criteria set forth for refunding of projects, six dealt with parent and PAC involvement.[4]

---

[2] D.P. Moynihan, Maximum Feasible Misunderstanding (New York: Free Press, 1969), p. 89.

[3] OE, Draft Follow Through Program Manual, February 24, 1969, p. 5.

Despite the continuing mandate for parental involvement in Follow Through, the attention given to parental involvement over the course of the national evaluation has waxed and waned. Initially, the evaluation gave this aspect of the program relatively little attention. However, after the October 1968 meeting in Atlanta and charges of insensitivity to parental and community involvement, SRI evaluators initiated a series of community studies exploring several models of the relationship between parental involvement and institutional change.[5] As the evaluation focused more narrowly on outcomes for children after 1970, attention to parental involvement clearly waned, and community studies were almost entirely dropped from the national evaluation.

Since then the main mechanism for evaluating parental involvement has been the parent interviews. These interviews have been conducted for Follow Through by the National Opinion Research Center (NORC), the data processed by SRI, and since 1972 analyzed mainly by Abt Associates

An early version of the parent interview form, largely open-ended, was developed and tried out in 1968. Over the next year the questionnaire was revised several times and long and short versions of the interview form pretested. After clearance of both versions of the form by the Bureau of the Budget in January 1970, the first wide-spread interviewing was begun in spring 1970. Altogether over the next six years, more than 60,000 parents in all four Follow Through "experimental" cohorts were interviewed, as is shown in Table 9.

---

[4] OE, "Follow Through Program," Federal Register, 39:121, June 21, 1974.

[5] SRI, Draft Appendix E: Follow Through Community Studies (Menlo Park, CA: August 1972), pp. E-8-16.

As is indicated in Table 9, the parent interviewing has been organized according to a sort of pre-post design. However, entry-year interviewing for cohorts I and II was not actually conducted at the time children entered Follow Through. Instead, it was "done late in the school year in order to give parents time to have had experience with the program and to develop attitudes toward it."[6] A small sample of parents was interviewed in fall 1970 in order "to obtain a measure before parents had any exposure to Follow Through", but large-scale fall interviewing did not begin until fall 1971 for cohort III. The switch was explained as follows:

> The overall design to date has called for interviews with parents during the first year their children are in Follow Through, followed by another interview with the same parents as their children "graduate" from Follow Through three or four years later. The change this year is to conduct the "entering" interview on a large scale as early as possible in the first year. This permits a longer time for changes to occur between the first and last interview. It also comes closer to "baseline" data--that is, the parents' attitudes, etc. prior to prolonged exposure to Follow Through. While it would be methodologically desirable to obtain a reading even earlier in the school year, experience has shown this is in fact virtually impossible.[7]

---

[6] OE, "Supporting Statement For OMB Clearance of the Spring 1972 Parent Interview Instrument," January 4, 1972, p. 10.

[7] OE, "Supporting Statement for OMB Clearance...," p. 10.

Table 11: Parent Interviews: Follow Through and Non-Follow Through Parents (and

Cohort-Stream			1969-70* Spring	1970-71** Spring	1971-72 Fall	1971-72 Spring	1972-73 Fall***
I	EK	FT	3814 (42)				
		NFT	2044 (40)				
	EF	FT	1923 (20)			660 (12)	
		NFT	1211 (20)			348 (12)	
II	EK	FT		5984 (66)			
		NFT		5077 (65)			
	EF	FT		2011 (22)			
		NFT		1992 (21)			
III	EK	FT			7570 (81)		
		NFT			6322 (80)		
	EF	FT			2557 (27)		
		NFT			2431 (25)		
IV	EK	FT					1158 (14)
		NFT					1121 (14)

Subtotals						
Cohorts	FT	5737	7995	10127	660	1158
I-IV:	NFT	3255	7069	7069	348	1121

Total Interviews		
Cohorts	FT	32,063
I-IV:	NFT	25,183
Other:		ca.6,100

*In addition to the Cohort I parents, roughly 5100 other parents were also
1400 implementation cohort parents and 3700 parents of fourth-grade students
Longitudinal Evaluation, pp. 28, 34.

**Slightly over 1000 parents were also interviewed in fall 1970 using the same
1970.

***Special interviewing of cohort II and III parents was carried out in fall
Roughly 1000 parents were specially interviewed for the summer study.

Sources: SRI, Follow Through Evaluation Data Summary (Menlo Park, CA: SRI,
SRI, Longitudinal Evaluation of Selected Features of the National
CA: SRI, March 1971).
J. David, Summer Study, Ed. D. dissertation, Harvard University,

Projects)

	1973-74	1974-75
Spring	Spring	Spring
950 (24)		
468 (24)		
	1631 (54)	
	1213 (52)	
504 (10)		
515 (10)		
		2808 (52)
		2903 (53)
	493 (14)	
	348 (13)	
1454	2124	2808
983	1561	2093

interviewed in early 1970: around
not in Follow Through. SRI,

instruments as were used in spring

1972 for the summer study.

April 1975).
Follow Through Program (Menlo Park,

1974.

Parent interviews were conducted only in sites included in the national evaluation. The sampling plan called for interviewing parents of all tested Follow Through children in each project unless there were more than 110 children, in which case 110 households were randomly selected for interviewing. For entry testing sponsors provided lists of parents with tested children directly to NORC.[8] Exit or third grade parents were, for the most part, interviewed only if they had been interviewed and their children tested when the children entered school.[9] Although virtually all Follow Through literature refers to this data-gathering as parent interviewing, the vast majority of interviews--around 95%--was with mothers.

Over the course of the five years of parent interviewing, several versions of the parent questionnaire have been used. The spring 1970 interviewing used two forms; a long form which took approximately an hour to administer and a short form which required only twenty minutes. In a manner reminiscent of what happened with the multiple forms of achievement tests for children, the two versions of the parent questionnaire were consolidated into a single form for the spring 1971 testing.

When the SRI parent interviewing was scrutinized by OE reviewers in summer 1971, they criticized several features of the effort, including

--the extensive changes in the instrument from year to year, and the consequent lack of continuity in data collected;

---

[8] SRI, Draft Appendix D, August 1972, p. C-90-91.

[9] A.R. Wright, Spring 1975 Parent Interview Cross-Tabular Findings (Menlo Park, CA: SRI, May 1976, revised November 1976), p. 1.

--the absence of any reliability data on the instrument.[1]

Despite the criticism of discontinuity in data collected, the parent interview form was revised again in 1972 and yet again in 1973. In spring 1971 the interview form had 70 questions, in fall 1971 63 questions, spring 1972 80 questions, in fall 1972 79 questions, and in spring 1973 and subsequent years 87 questions.[2] The 87-question version, used for three years in a row, contained more than 200 subquestions altogether. Many of the questions and subquestions in the 87-question version were the same as ones in earlier versions, but some questions were changed in every version prior to 1973.

The continual revising of the interview form up until 1973 reflected a discontinuity deeper than form; and this was one that never completely stabilized. Exactly why were the parent interviews being conducted? And how were data from them to be used? Describing the first round of parent interviewing in spring 1970, SRI evaluators wrote that the purpose of the interviews

---

[1] Judy Crooks and Donald Burnes, memorandum "Consolidated Review of September 17 Deliverables" to Fred Bresnick, Acting Chief, Research and Evaluation Section, Follow Through Program, October 6, 1971.

[2] SRI, *Follow Through Evaluation Data Bank File Description All Years Beginning 1972-73* (Menlo Park, CA: revised January 1975) and preceding data bank files. Counting how many questions were in each interview form is a bit difficult because it depends on what you call a question. Many questions had several parts and followup questions depending on previous responses. The numbers given here refer to questions identified as such in the SRI *Data Bank File Description* volumes and do not necessarily correspond to the numbering of questions on the actual interview forms.

> was to identify certain attitudes and actions of parents and to obtain information about demographic and other characteristics of the household. The survey concentrated on collecting data that might be related to the development and educational progress of the child, might change over time as a consequence of participation in Follow Through, and would be helpful in estimating the comparability of Follow Through and non-Follow Through children.[3]

Elsewhere in the same report the SRI evaluators also suggested that the parent responses might be used as indicators of the degree of implementation of Follow Through projects.[4]

In their 1973 report the SRI investigators expanded on four main purposes of the parent interviews:

> (1) Provide information from which to estimate the initial comparability of Follow Through and non-Follow Through children and families according to socioeconomic, ethnic, and other demographic characteristics, and to adjust statistically for non-comparability since random assignment was not possible.
>
> (2) Provide information for sorting respondents into subgroups so that possible interactions between treatment and parent characteristics can be examined.
>
> (3) Develop indicators of parental beliefs, expectations, and practices that characterize family life styles which may be influenced by or may mediate the effects of Follow Through program participation. Some Follow Through programs are specifically designed to bring about a considerable degree of parent involvement and parent education while others emphasize these goals only minimally.
>
> (4) Ascertain the parent's knowledge of, participation in, and satisfaction with school programs generally and Follow Through in particular.[5]

---

[3] SRI, *Draft Longitudinal Evaluation of Selected Features of the National Follow Through Program* (Menlo Park, CA: SRI, March 1971), p. 27.

[4] SRI, *Draft Longitudinal Evaluation*, p. 52.

[5] Emrick et al., *Interim Evaluation*, p. 40.

This reasoning was not compelling to the review panel instituted after the Guttman and Willner charges against the evaluation. The panel recommended that "that parent questionnaire...should be included in local evaluation efforts and not in the national evaluation as in the past."[6] The recommendation apparently had little impact on the evaluation, probably because by that time work on the parent interview data was well under way. By 1973, several million dollars had already been spent to gather the parent data. The spring 1972 RFP for analysis of national evaluation data had already set out as a principal issue the question "Does the Follow Through program in the aggregate have a greater impact on parents and children than non-Follow Through schooling?"[7] And the original Abt proposal had naturally responded to the OE request:

> [C]hanges in attitudes and behavior in both teachers and parents should constitute domains of dependent variables.... It is also clear that teacher and parent measures shall also serve as both independent and control variables in the examination of child scores and in identifying sponsor effects.[8]

The two uses for the parent interview data mentioned in the Abt proposal--as dependent variables or as independent and control variables--have not come to equal fruition. Throughout all of the major outcome evaluations, that is, the Emrick and the Abt I-IV reports, at least some data from the parent interview questionnaires have been used as covariates to adjust for initial differences between FT-treatment groups and NFT-control groups. Most commonly used have been the data on mother's education and family income. although the Abt II report explored a wider range of parent data covariates.

---

[6] Second Report of the Follow Through Evaluation Review Panel, March 1973, p. 11.

[7] OE, Request for Proposal Follow Through, March 13, 1972, p. 2.

[8] Abt, Proposal May 5, 1972, p. 9.

In the second realm--the use of parent data to evaluate program outcomes for parents--approaches have proved considerably more varied, as shown in Table 12. In part this is due to the lack of any clear-cut plan as to how the parent data were to be analyzed. Although the parent interviewing was organized in pre-post fashion, apparently no plan was ever formulated for exactly how it would be analyzed. As late as 1972, it was disclaimed

> It is inappropriate to try to specify in detail at this time the precise analyses that will be used in the Parent Interview since the Parent Interview is an integrated part of the overall Follow Through evaluation.[9]

Another problem confronting every would-be analyst of the parent data is simply that there is so much of it--approximately 200 pieces of information for each of 60,000 interviews, or roughly 12,000,000 overall. The amount of parent information analyzed in any one of the Follow Through evaluation reports is substantially less than this total--probably only one or two million--but the first task for everyone approaching the parent data still has been simply how to reduce it. Emrick and his colleagues did so by selecting "judgmental groupings of criterion items on the PI schedule..."[1] The SRI evaluators used this method to develop seven different scales of parent outcomes, although only four were reported in their 1969-71 report.

---

[9] OE, "Supporting Statement for OMB...," January 4, 1972, p. 15.

[1] Emrick, p. 61.

The Abt evaluators used three different methods for handling the parent interview data: judgmental selection, cluster analysis, and factor analysis.[2] These approaches have yielded quite a variety of synthesized variables as shown in Table 12. Successively the Abt evaluators have derived six clusters for the cohort III kindergarten parent data, eight clusters for cohort I third grade data, ten factors for cohort II third grade data, and thirteen factors in the final Abt IV analyses. As is suggested by their names, the synthesized variables in the different reports are often similar, sometimes containing identical questions. Yet more often than not, they also contain different questions, so this has helped frustrate the analysis of the parent data in pre-post fashion--the design according to which the data were gathered.[3] So the analyses of the parent data have fallen somewhat short of McDaniels' supposition that the existing interviews would "assess changes in parents and their reactions to the program."[4]

---

[2] Factor and cluster analysis are techniques for analyzing patterns of variation among variables and empirically combining or grouping together those variables which covary. Ideally the result is that a large number of variables can be reduced to only a few major dimensions.

[3] The Abt III report mentioned that "Due to the late administration of the entry Parent Interview, only the exit interview [for cohort II] was selected for use in this analysis" (p. A-25). However, this could not have been a reason for not analyzing the C III entry parent data, as even the Abt III report points out.

[4] Garry McDaniels, "Design of the Follow Through Experiment" in Rivlin and Timpane, Planned Variation in Education (Washington, D.C.: Brookings Institution, 1975) pp. 59-60.

Table 12: Analyses of Parent Interview Data

REPORT:	Emrick, Interim Evaluation	Cline, Abt I		Stebbins, Abt III	Stebbins, Abt IV
COHORT STREAM, GRADES & GRADE, YEAR DATA GATHERED:	CI EK, SK, 1970; CI EF, S1, 1970; CII EK, SK, 1971; CII EF, S1, 1971	CIII EK, FK, 1971.	CI EF, S3, 1972.	CII EK, S3, 1974.	CII, EF, S3, 1973; CII EK, 1974; CIII EF, S3, 1974; CIII EK, S3, 1975.
VARIABLES DERIVED FROM PARENT INTERVIEW DATA:	Four scales - Parent-child interactions (11,4) - Parent-school involvement (5,5) - Child academic expectations (3,3) - Parent sense of control (6,4) (Three other scales were calculated but were not included in the SRI interim analysis "generally because of an excessive amount of missing data or inadequate response variability.") - Parent-social interactions - Parent locus-of-control - Parent participation in PAC	Six clusters - School's receptivity to parent (3) - Parent-school interaction-A (4) - Parent-school interaction-B (3) - Parent: child's school-oriented behavior (4) - Parent satisfaction with child's academic progress (3) - Parent locus-of-control (3)	Eight clusters - Knowledge of FT (4) - School's receptivity to parent (3) - Parent-school interaction (4) - Parent-school interaction (3) - Parent: child's school-oriented behavior (6) - Parent satisfaction with child's academic success (3) - Parent satisfaction with child's affective growth (4) - Parent locus-of-control (3)	Ten factors - School helped child (11) - School helped parent understand (8) - Child's progress in school (6) - PAC involvement (3) - Parent visits school (4) - Parent-child interaction (6) - Parent locus-of-control (3) - School's receptivity (3) - School helped parent personally (3) - Satisfaction with teacher (3) (Two other factors were calculated but were not reported "due to lack of interpretability in terms of sponsor goals and objectives.") - Child's emotional adjustment (3) - Talks with teacher (3)	Thirteen factors " combined into four order factors" I. SCHOOL IMPACT ON CHILD - School helped child (10) - Parent satisfied with progress in school (5) - Child's emotional ment (4) - School helped parent child (8) - School's receptivity II. PARENT ACTIVITY - PAC involvement (4) - Parent visits school - School helped parent personally (3) - Talks with teacher III. PARENT-CHILD - Child reads at home - Child read to at home - Parent-child IV. PARENT LOCUS - Parent locus-of-control
SOURCE:	Emrick, et al., Interim Evaluation, February 1971, pp. 61-62. How many questions are in each of the scales is not clear in the Emrick report. The numbers in parentheses above, for the 1970 and 1971 survey data respectively are from SRI, Appendix C of Draft Evaluation of the National Follow Through Program, Menlo Park, CA: SRI, August 1972. Also, pp. 11-18, 40.	Cline et al., March 1974, Abt IB, pp. 1-5-6.	Cline et al., Abt IB, March 1974, pp. 1-33-34.	Stebbins, et al., Abt IIIA, July 1976, pp. A-28-29.	Stebbins, et al., 1977, pp. 31-34.

*In parentheses after each scale, cluster or factor are the number of questions or sub-questions combined together to make each one.

The continuing revisions in the variables derived from the parent interview data are partially due to the changes in questions administered in different interviews. Also however, it may derive from what Edward Cureton once called one of the "Principal Compulsions of Factor Analysts." Cureton observed that people who indulge in factor analysis are likely to believe all other factor analysts insane and are highly prone to criticism and re-interpretation of results obtained by methods other than their own.[5]

Another cause for reinterpretation of the parent interview variables has been pressure from Follow Through sponsors. When Abt evaluators preparing their third report described variables derived from parent questionnaires as empirical estimates of children's home learning environment, several sponsors objected. One maintained that "parents' perceptions of the program are so personalized and impressionistic that they cannot be termed empirical estimates."[6] Another complained

> Some of the items in the "satisfaction" factor from the parent questionnaire seem beyond the influence of the Follow Through program in most field settings (particularly items on "child's progress in school" and "satisfaction with teacher"). Yet, when parent questionnaire findings are reported, it is implied that the sponsor either succeeded or failed to achieve parent satisfaction.[7]

---

[5] E.E. Cureton, "The Principal Compulsions of Factor-Analysts," Harvard Educational Review 9:3 (May) 1939, 287-295. Observation suggests that the compulsions of factor analysts covary very highly with those of cluster analysts.

[6] Elizabeth Gilkeson, letter to Director Follow Through Evaluation, Abt Associates, February 18, 1976, enclosure.

[7] J.T. McNeil and J.T. Bond, "Technical Comments in Response to Abt Follow Through Report", High/Scope Educational Research Foundation. February 24, 1976.

As a result, the final version of the Abt III report called the parent variables (and those derived from the teacher questionnaire as well) "indications of acceptance of the model tenets."[8] The linguistic devaluation of the parent and teacher data--from empirical estimates of implementation to indications of acceptance--carried over into the Abt IV report. The parent and teacher data were not even mentioned in the final "patterns of effects" chapter in the main volume of the Abt IV report nor in the digest of the Abt IV report.

The fate of the parent and teacher data--partially as a result of sponsors' advice--was somewhat ironic. For the parent and teacher data contained information which was considerably more complimentary to Follow Through as a whole than any of the student test data. The test data on which the national evaluation finally focused suggested the effects of Follow Through as a whole to have been more negative than positive. In contrast, however, when parents of cohort III third grade children were asked how much school had helped them find medical, dental, and other services when their children needed them, 47% of Follow Through parents said "a great deal" as compared with only 17% of non-Follow Through parents.[9]

---

[8] Abt III, p. A-24.

[9] SRI, Feedback Data for Spring 1975, Parent Survey Data, p. 77.

Evaluation of Teacher and Teacher Aide Questionnaire Data

> The financial and verbal commitment of a school district to a model did not mean that everyone, least of all teachers, needed only to be told about the model in order to implement it. There were problems both of motivating teachers to change their teaching styles and of supporting orderly change in the classrooms rather than expecting the revolution to come overnight. The early workshops that were expected to stimulate teachers to create open, active classroom environments where before there had been teacher domination and reliance on highly abstract materials were well received. The new ways of doing things were enthusiastically accepted; but a few months later, change was difficult to detect in the classrooms.
>
> -David Weikart and Bernard Banet[1]

The role of teachers in implementing a new educational program in schools might seem obvious to outsiders, but teachers and teacher aides often seem to have been the forgotten people of Follow Through. The 1969 program manual devoted several pages to parent and community involvement in Follow Through, but gave only passing reference to teachers and teachers' aides. In describing project selection of program models, the manual stated only that "All persons who will participate in the local Follow Through project--parents, other community representatives, project staff--should consider this information [on a range of program approaches] and take part in the discussions leading to the selection of an approach."[2] Despite little discussion of what teachers' role would be, the manual envisioned that it would be a demanding one:

---

[1] D. Weikart and B. Banet, "Model Design Problems in Follow Through" in Rivlin and Timpane, pp. 63-64.

[2] Office of Education, Draft Follow Through Program Manual, February 24, 1969, p. 11.

> Follow Through will make new demands on participating teachers, especially in initial implementation stages. Therefore, regular teachers of Follow Through children should not teach more than one class (group) of children. Other non-project responsibilities should be reduced to a minimum.[3]

In the early phases of the national evaluation, teachers were included in two ways. First, they were observed and their behaviors recorded and analyzed; by Jane Stallings and her group at SRI, by Robert Soar and others at the University of Florida, and by Joseph Grannis and his colleagues at Teachers College. The classroom observation studies aimed mainly at assessing implementation of Follow Through models. Soar and Stallings also explored the relationship between classroom variables and student test data, as discussed in chapter III. Final reports from the Soar and Grannis studies were issued in 1973 and the final Stallings report in 1974.

Since then, the sole source of information on teachers and teacher aides, and the only one drawn upon in the Abt outcome evaluations, has been the teacher and teacher aide questionnaires. Like the parent interviews, the Follow Through teacher questionnaire has had an up and down career. As mentioned in the last section, the career of the teacher questionnaire in the national evaluation has paralleled that of the parent interview in many instances. In several respects, however, the teacher and teacher aide questionnaires have not paralleled the parent interview.

---

[3] Office of Education, Draft Program Manual, p. 17.

The major difference between the parent interview and the teacher and teacher aide questionnaires is in the form of administration. Unlike the parent interview, the teacher questionnaires were administered not in person, but by mail. Work on a teacher questionnaire began in the summer of 1968, shortly after SRI won their initial Follow Through evaluation contract. As with the parent interviews, widespread use of the teacher questionnaire did not begin until 1969-70. In May 1970, questionnaires were mailed to all Follow Through coordinators for distribution to all Follow Through and Non-Follow Through teachers whose children had been tested in fall 1969 or spring 1970.[4]

A similar procedure was followed the next year--except that while the two versions of the parent interview form were undergoing fusion, the teacher questionnaire underwent fission. A single version of the teacher questionnaire was used in spring 1970, but two were created for the spring 1971 administration--one grey and one salmon.

> Two versions were developed since the full content of questions would have created a version thought to be excessively long. Thus, each version omitted certain items that were contained in the other; both versions contained a major fraction of overlapping content.[5]

When the Office of Education officials reviewed the Follow Through evaluation in the summer of 1971, the teacher questionnaire, like the parent interview form, drew criticism. The OE reviewers "expressed concern about the extent of change in the items from year to year, and the consequent lack of continuity in the data.'[6] They recommended that SRI subcontract the development of the teacher questionnaire to a mutually agreed upon subcontractor. As a result, an initial version of the 1971-72 questionnaire was developed by a group at Stanford University, but it was "modified extensively by USOE/FT and their advisors"--including McDaniels' consultants at Huron. So teacher and aide questionnaires changed almost as often as the parent interview forms. According to the SRI data bank file description, the number of questions in the different versions of the questionnaires were as follows:

---

[4] SRI, *Draft Appendix C Follow Through Pupil Tests*..., August 1972, p. C-107.

[5] SRI, *Draft Evaluation of the National Follow Through Program 1969-71*, p. II-24.

	1969-70	1970-71	1971-72	1972-73 and subsequent years
Teacher questionnaire	40	46	49	73
Aide questionnaire	19	32	24	41

There was also a change in the manner in which teacher questionnaires were administered in 1972. In the first two years of the teacher questionnaire, the SRI investigators had been troubled by a low return rate on the questionnaires. In spring 1971, they sent out around 2800 teacher and aide questionnaires, but despite follow-up letters and phone calls, achieved a response rate of only 80%.[7] The follow-up effort not only proved costly, but also prompted complaints from teachers. These problems were largely overcome in spring 1972 by "awarding a small honorarium to each teacher and aide who completed a questionnaire." The $10 honorarium helped prompt a response rate of 97 percent in spring 1972.[8]

---

[6] J. Crooks and D. Burns, "Consolidated Review," October 6, 1971, p. 10.

[7] Draft Appendix C Follow Through Pupil Tests..., August 1972, pp. C-108-08.

Such issues of procedure seem to have been more tractable than those of purpose. If anything, the exact purposes and plans for analyzing the teacher questionanire data seem to have been more ambiguous than those concerning the parent questionnaire. The Office of Education review of the evaluation in September 1971 had noted that "The rationale for collecting data on all aspects of program process description [including the teacher questionanire] was inadequate."[9] Yet curiously the reviewers did not call for delineation of the rationale.

A request for clearance of the 1972 teacher questionnaire form provided the following rationale:

> The Follow Through program will attain successful outcomes by influencing teachers to utilize innovative classroom practices in a consistent and competent manner. In addition, this program might improve the teachers' basic knowledge concerning which types of educational methods are most effective, and which principles of learning and teaching can be most validly utilized with disadvantaged children. Furthermore, significant increases in the Follow Through teachers' sense of competence in attaining the program's objectives might also occur. Consequently, it is essential to administer a questionnaire to Follow Through and "comparison" group teachers in order to gather data concerning the differences that may exist between these groups in their educational knowledge and basic attitudes toward teaching. The present questionnaire has been designed to obtain these data.[1]

---

[8] SRI, Draft Evaluation of the National Follow Through Program 1969-71, August 1972, p. II-24.

[9] J. Crooks and D. Burnes, "Consolidated Review...," October 6, 1971, p. 9.

[1] OE, "Supporting Statement for OMB Clearance of the Spring 1972 Teacher Questionnaire," n.d., p. 1.

The Follow Through review panel recommended in 1973 that the teacher questionnaire continue to be used in the national evaluation, but gave no elaboration as to why. The Emrick report explained that the teacher and aide questionnaires,

> were developed to complement and support data gathered through other measurement instruments of the Follow Through evaluation and to learn whether teachers in the Follow Through program were changing. First, the questionnaires were designed to provide profiles of the teachers in terms of demographic characteristics, the training and support they received, and the goals and attitudes they held. Second, the questionnaires could provide information about the different effects of different Follow Through sponsors' programs.[2]

Describing analyses of the teacher questionnaire data, the Abt I report was similarly vague:

> While we have not yet merged teacher and pupil data, the teacher studies shed some light on variations in a number of important teacher characteristics including: (1) personal and professional background; (2) training in basic Sponsor philosophy; (3) values, attitudes, and reported behaviors; and (4) satisfaction and perceived faithfulness to the Sponsor's approach. The teacher studies also explore the relationship of teacher background to the other teacher characteristics, in order to determine whether or not Sponsor delivery and/or implementation are mediated by the characteristics teachers bring with them to the program.[3]

The Abt II report proclaimed, "it is essential that meaningful descriptors of the teachers be included" in the national evaluation because

---

[2] Emrick, Interim Evaluation, pp. 41-42.

[3] Cline, Abt IA, March 1974, p. VI-3.

> Teachers mediate the Sponsor's model; they tend to set the tone of the experiment as it impinges on children; and they contribute to the decisions which determine the formal shape of the program. We assume that teachers therefore influence pupil outcomes to varying degrees. Hence the teacher is one of the critical elements of the educational program which is delivered to each child.[4]

In the Abt III report, the teacher data underwent the same small linguistic devaluation as the parent interview data. Initially, the variables derived from the teacher questionnaire were introduced as "indications of implementation of the learning environment", but after sponsors objected were devalued to "indications of acceptance of the model tenets". In the Abt IV report the teacher data retained the same status, but like the parent data faded into the text of the report and were not even mentioned in either the final chapter on patterns of effects nor in the digest of the report.

On the way to this relative oblivion, the teacher questionnaire data have undergone a variety of analyses as shown in Table 13. In the Emrick study, items from the teacher questionnaire "were assembled into logical and judgmental groupings, responses were rescaled after dichotomizing the distribution for each item at or nearest the median; the rescaled item scores were then aggregated into a grouping or variable score", and then the scaled variable scores were transformed into a standardized metric.[5] In the Abt analyses, the teacher data underwent various kinds of cluster analysis. But while the teacher data underwent only the compulsions of cluster analysts and not those of factor analysts, the results seem similar. The number of

---

[4] Cline, Abt IIB, May 1, 1975, p. II-1.

[5] Emrick, Interim Evaluation, February 1973, p. 63. The Emrick study used these same kinds of transformations to standardize the parent interview data.

Table 13: Analyses of Teacher

REPORT:	Emrick, Interim Evaluation	Cline, Abt I
COHORT, STREAM, SEASON & GRADE, YEAR DATA GATHERED:	CI EK S1, 1971; CI EF S2, 1971; CII EK SK, 1971; CII EF S1, 1971.	Spring 1972, kindergarten and other grades teachers.
VARIABLES DERIVED FROM TEACHER QUESTIONNAIRE DATA:	Two scales - Parent-educator image (the extent to which teachers reported they felt it essential to "get together with parents outside of the classroom") - Professional acceptance of method (the extent to which the teacher reports she would not prefer to adopt some teaching approach other than the one she is using)	Nine clusters FT ONLY - Teacher satisfaction (3) - Perceived faithfulness to sponsor's approach (6) - Amount of training: structure (3) - Amount of training: child centeredness (8) - Amount of training: working with parents & aides (3) FT AND NFT - Teacher attitudes toward meeting with parents (6) - Teacher values: parent-community orientation (3) - Teacher values: social skills development (3) - Teacher values structured academic vs. child-centered orientation (10)
SOURCE(S):	Emrick, Interim Evaluation, February 1973. Also, Appendix C of the August 1972, SRI Draft Evaluation.	Cline, Abt IB, March 1, 1974, pp. II-17-18

Questionnaire Data.

Cline, Abt II	Stebbins, Abt II	Stebbins, Abt IV
Spring 1973, CIII SK, first grade teachers.	Spring 1974, third grade teachers.	Spring 1974 and Spring 1975, third grade teachers.
Fourteen clusters	Five clusters	Five clusters
FT ONLY	FT ONLY	FT ONLY
- Teacher satisfaction with FT sponsor (6) - Amount of PAC activity (6)	- Satisfaction with sponsor (6)	- Satisfaction with sponsor (6)
FT AND NFT	FT AND NFT	FT AND NFT
- Amount of teacher training: social/emotional/intellectual development & individualization of instruction (7) - Amount of teacher training: classroom management (5) - Amount of teacher training: parent involvement (2) - Reported teacher practices: emphasis on basic skills vs. affective development (3) - Reported teacher practices: flexible/open classroom (4) - Reported teacher practices: individualization of instruction (3) - Reported teacher practices: teacher centeredness (4) - Importance of parent involvement: social/educational activities (10) - Importance of parent involvement: decision-making activities (3) - Benefit of parent involvement (4) - Teacher satisfaction with pupils' affective development (4)	- Teacher centeredness (4) - Flexible classroom (4) - Individualization of instruction (3) - Academic vs. emotional emphasis (3)	- Teacher centeredness (4) - Classroom flexibility (4 - Individualization of instruction (3) - Academic vs. emotional emphasis (3)
Cline, Abt IIB, May 1, 1975, pp. II-9-13.	Stebbins, Abt IIIA, July 1976, pp. A-25-26.	Stebbins, Abt IVA, February 1977, pp. 29-30.

clusters derived from the teacher data has ranged widely, from nine to fourteen to five.

Initially teacher data were used as covariates in analyses of student test data as well as being used as measures of program implementation. For example, Emrick used eight kinds of data from the teacher questionnaire as "input data on teachers," and hence as covariates in analyzing pupil test data.[6] Variables from the teacher questionnaire, like teacher's education, teaching experience, and teacher's ethnicity were also used as covariates in some of the Abt I analyses. However, the teacher questionnaire variables were not used as covariates in any of the subsequent Abt analyses. The disappearance of the teacher data as "input/control" variables was undoubtedly related to the shifts in level of analyses. Both the Emrick and the Abt I studies included analyses using the classroom as the unit-of-analysis and some teacher variables correlated significantly with classroom average scores. Later analyses (in the Abt II, III and IV reports) used the child as the unit-of-analysis and teacher covariates seem to have been dropped in these analyses at least partially because teacher variables do not explain much of the variance in individual children's test scores.

Use of the teacher-aide questionnaire data has not varied as has that of the teacher data. The reason is quite simple. Although the teacher aide data have been used in SRI feedback tables to projects, these data have not been used in any of the major national evaluation studies.[7] As Gilbert Krulee observed several years ago,

---

[6] Emrick, Interim Evaluation, pp. 75-89.

> The danger for projects like Follow Through is not that they will collect too little data but rather that they will find themselves overwhelmed with a volume of data that they cannot possibly handle. Then the "message" becomes lost in a background of irrelevant noise.[8]

Not just the "message," but apparently also some of the data got lost in the sea of data.

Summing Up

By the spring of 1977, the parent, teacher and teacher aide data had almost completely faded from view in the national Follow Through evaluation. Neither parent or teacher data were mentioned in the digest or summary of the Abt IV report. These data did, however, get used in two sorts of reports outside the mainstream of the evaluation--in SRI feedback tables to Follow Through projects and in two "cross-tabular" reports prepared by SRI in 1976. So before summing up the fate of the parent and teacher data in the national evaluation, let me briefly describe these uses of the parent and teacher data.

How to make Follow Through national evaluation data available to local Follow Through projects became an issue quite early in the history of Follow Through. As Gilbert Krulee put it, "There was an expectation that SRI would 'give' as well as 'take,'...[and] would make available processed data that would be helpful to projects and sponsors."[9] Early in the 1969-70 school year selected data from the 1968-69 data-gathering were returned to local Follow Through project personnel, but since "this feedback consisted primarily of uninterpreted descriptive tables," it was "apparently of only limited utility for local planning."[1] In May 1970, the Goldberg review panel recommended that "mechanisms for providing informational feedback from SRI should be more clearly defined and implemented."[2]

---

[7] Whether any of the teacher aide data were actually used in analyses reported in the Emrick report is unclear. Introductory discussion (pp. 52, 58) indicates that data from aide questionnaires were used in analyses, but reported results (pp. 82-89) do not seem to emcompass any of the teacher aide data.

[8] G. Krulee, Organizational Analysis, April 1, 1973, pp. 208-09.

Apparently, however, little progress was made on the feedback issue until 1972. OE policies for "providing information resulting from the national evaluation to participating communities, sponsors, State Education Agencies and State Office of Economic Opportunity" were originally set out in a memorandum in March 1972. Among these policies were the following:

- Information will be made available to various audiences as promptly as possible for the purposes of program revision and improvement.
- All interpretations of National Evaluation data released to public news media must clearly state the source of the interpretations.

---

[9] Krulee, Organizational Analysis, p. 212.

[1] Krulee, Organizational Analysis, p. 212.

[2] SRI, Draft Administrative History, p. F-22.

- Data collected from individuals with assurances of individual anonymity will not be reported in such a way as to betray the confidence of those individuals.[3]

Feedback information from 1969-71 data gathering was sent out by SRI in June 1972. Apparently these feedback data were not widely circulated, however, because one of the questions put to OE officials by parent representatives at the October 1972 Denver conference was "Why aren't data being sent back to parents?" In response to the questions raised at the Denver meeting, the official in charge of the Follow Through evaluation explained:

> Certain pupil, teacher, and parent data collected for the national evaluation are sent to all local PAC chairmen, Follow Through directors, and school superintendents according to the national evaluation feedback procedures established by the U.S. Office of Education (USOE). The inclusion of PAC chairmen reflects USOE's concern for reaching Follow Through parents...
>
> In the initial stages of the national evaluation, USOE was concerned about the premature release of data. This concern was based on the realization that such information, if misinterpreted, could prejudice programs before they could be established as properly implemented projects...
>
> Since the program has matured, plans now call for much more evaluation data to be made available to local communities.[4]

---

[3] OE, Memorandum "U.S. Office of Education Policy for National Evaluation of Follow Through," March 1972, p. 3.

[4] Edward Chalker, Acting Chief of Research and Evaluation, "Memorandum to PAC Chairmen, Members of the PAC Evaluation Committee, and Follow Through Directors. Subject: Questions Raised at the Denver Meeting, 1972," April 6, 1973. Chalker was acting chief of the Follow Through evaluation in the interregnum between McDaniels' departure and Tucker's appointment as head of the evaluation.

In accordance with this policy, feedback data were regularly distributed to local officials involved with Follow Through (and to USOE) over the next three years. Data gathered in the 1971-72 school year were sent out in June 1973. Data collected in spring 1973 were mailed out during the first half of 1974, and data from spring 1974 and spring 1975 were distributed in early 1975 and 1976 respectively.

Reports of feedback data had as many as four parts:

- Roster Data
- Pupil Test Data
- Teacher and Aide Survey Data
- Parent Survey Data.

Roster data were returned to almost all Follow Through projects. Test, parent and teacher data were sent back only to those projects in which these respective types of data had been gathered. In each feedback document data were reported for the local Follow Through group overall, and as comparisons, for the overall national Follow Through sample, and the overall national Non-Follow Through sample. In cases where all four types of data were available, the feedback reports were impressively weighty--around 300 pages for each year of data gathering. Each feedback report was accompanied by a manual recounting the limitations of the aggregate data:

> It should be noted that data distributed as feedback from the national evaluation data collection are unanalyzed descriptive data. Any analysis, interpretation, or inferences drawn from these data by feedback recipients will be the responsibility of such recipients. All interpretations of national evaluation data released to the public must clearly state the source of the interpretation.[5]

No information has been systematically gathered on how or if Follow Through projects made use of this data, but SRI staff report that the feedback data were "well received by people in the field."[6] At least some projects did make use of the feedback data in their own evaluation activities. For example, in a presentation to the House Appropriations Subcommittee on the Departments of Labor and Health, Education and Welfare in 1973, the director of the Portland, Oregon Follow Through project, explicitly cited the feedback data. Recounting data from the parent survey--for example, that 55 percent of the Portland Follow Through parents surveyed had visited their children's classrooms compared with only 33 percent of Non-Follow Through parents--he concluded that "parental involvement in school activities is one of the most beneficial aspects of the Portland Follow Through program."[7]

---

[5] SRI, Follow Through Feedback Manual (Menlo Park, CA: SRI, February 1975), p. iii.

[6] Teresa Middleton, SRI, letter to Walter Haney, Huron Institute, February 18, 1977.

[7] Mr. Robert Harold, statement to the House Appropriations Subcommittee on the Departments of Labor and Health, Education and Welfare. Hearings February 21, 1973.

The other use of parent and teacher data outside of major evaluation reports was in cross-tabulations of results from the spring 1975 parent interviews and teaching staff questionnaires. The cross-tabulations of 1975 parent and teacher data had originally been prepared by SRI as part of an effort to plan for a "continuing effects" study of Follow Through children beyond the third grade. The idea was that Follow Through parents might be interviewed in later years after their children had left the Follow Through program, and so SRI tabulated results from the 1975 interviews in order to explore how interview procedures might be refined in the follow-up study.[8] Plans for the continuing effects study fell through, but OE nevertheless requested that SRI write up its cross-tabulations of the 1975 parent and teacher data as separate reports.[9]

The SRI parent and teacher interview reports compared responses mainly at the overall Follow Through versus Non-Follow Through level. Selected items were discussed at the sponsor level although data were not systematically reported for different sponsors. At the aggregate levels, the cross-tabular findings revealed several interesting findings. In both parent and teacher surveys, responses indicated that Follow Through children were receiving more adequate medical and dental services than Non-Follow Through children.[1] The parent survey also suggested that "Follow Through parents, more often than Non-Follow Through parents indicate that school has been helpful to their children and to themselves."[2]

---

[8] One change in procedures that was being considered was to conduct interviews via telephone rather than in person. In-person interviews had cost in the neighborhood of $50-$80 each in 1975, and the hope was that costs could be substantially reduced by switching to telephone interviews. Results of SRI pretesting of telephone interview procedures are reported in Teresa Middleton and Anne Wright, Report on Parent Interview Pretest Findings (Menlo Park, CA: SRI, August 1976).

[9] Anne R. Wright, Spring 1975 Parent Interview Cross Tabular Findings (Menlo Park, CA: SRI, May 1976), 144 pp. Revised version: November 1976, 150 pp. Connie L. Sutherland, Teresa Middleton and Anne R. Wright, Spring 1975 Teaching Staff Questionnaire Cross Tabular Findings (Menlo Park, CA: SRI, May 1976), 176 pp.

Such findings from the cross-tabulations of parent and teacher data were of obvious pertinence to the broad goals of Follow Through as a whole, but they were completely out of synchronization with Abt's analyses of test data. By 1976, Abt had completely dropped overall Follow Through versus Non-Follow Through comparisons and aimed their analyses at estimating effects at the site level. So there was no way in which the results of the Abt III and IV reports could be directly compared with SRI's cross-tabulations of parent and teacher data. The two efforts were aimed at completely different levels of aggregation.

This problem of coordination is doubtless one of the main reasons why the parent and teacher data faded from view in the national evaluation. The main focus of the national evaluation was on children's test data--after all, the ultimate aim of Follow Through is to aid in the continued development of disadvantaged children to their full potential, and more immediately, to amplify the educational gains made by such children in Head Start. So as attention in the national evaluation focused on children's educational gains, analyses of parent and teacher data were conducted as separate endeavors, with little attention given to how they might be coordinated with analyses of children's test data.

---

[1] Wright, Parent Findings, p. 51. Sutherland, et al., Teacher Findings, pp. 69-70.

[2] Wright, Parent Findings, pp. 2, 46, 52.

Another reason for the fading from view of the parent and teacher data was that it did not fit well with the prevailing paradigm for program evaluation. Although in the early years of Follow Through numerous people spoke of program effects on parents and teachers, no one seems ever to have formulated a plan for exactly how interview and questionnaire data would be used to infer program effects on parents and teachers. And even if there had been such a plan, it was evident by 1976 that any analysis of parent and teacher data in pre-post fashion would have to deal with several formidable problems. For example, the 1975 teacher survey indicated that teachers switched into and out of the Follow Through program just about as fast as students. Of teachers responding to the 1975 survey, less than 40 percent reported more than three years' experience with the Follow Through approach they were currently using.[3] Apparently as a result, no attempt has been made to link teacher questionnaire data for particular teachers across more than one time point.

One attempt was made, however, to link parent data in pre-post fashion. In 1976, SRI conducted an exploratory analysis of six items which were included in at least similar form on both the fall 1971 entry and spring 1975 exit parent interviews.[4] The analysis was conducted on data from 3,455 parents who had been interviewed in both fall 1971 and spring 1975. Several problems emerged in the exploratory analysis. First of all, most of the interview questions were worded slightly differently in the 1975 interviews than in the 1971 ones. Second, as the SRI authors put it, "parent responses tend to be volatile."[5] Although there were some net changes in the way the traced sample of parents responded to particular questions between 1971 and 1975, the net changes generally were smaller than raw changes in opposite directions. For example, when asked whether they agreed or disagreed with the statement, "People in our schools really care about what parents think," 77 percent of the traced sample of Follow Through parents said they agreed in 1971. When asked a similar question in 1975, 81 percent of the same sample said they agreed; for a net shift of 4 percent. However, 10 percent of parents who said they agreed in 1971 said they disagreed in 1975 and thirteen percent of those who disagreed with the statement in 1971 said they agreed with it in 1975.[6] Such volatility--or unreliability as it might also be called--makes it clear that inferring that Follow Through caused any such shifts in parent responses can be done only with a very low level of certainty. Also, it suggests that the recommendation made by OE reviewers in 1971--that a reinterview study of the reliability of the parent interview instrument should be conducted--was a wise one.[7] Apparently, however, despite the fact that more than 60,000 parent interviews were conducted for the national Follow Through evaluation, no such interview-reinterview study was ever carried out.

---

[3] SRI, Duluth, Minnesota Feedback Data for Spring 1975 Teacher Survey (Menlo Park, CA: SRI, 1975), p. 119. The high rate of teacher mobility into and out of Follow Through suggested by this figure also has several implications for the implementation issue--particularly for sponsors who report that it takes several years of experience for a new teacher to become an effective practitioner of their approach.

---

[4] Kaye Crawford and Teresa Middleton, "Comparison of Responses to Selected Items Contained in Fall 1971 and Spring 1975 Parent Interviews," Menlo Park, CA: SRI, June 1976.

[5] Crawford and Middleton, "Comparison of Responses," p. 3.

[6] Crawford and Middleton, "Comparison of Responses," p. 6. The pattern evident in overall Follow Through comparisons was also evident in comparisons of responses examined on a sponsor-by-sponsor basis--that is, net changes in responses tended to be smaller than raw changes in both directions. The data were not disaggregated to the site level.

So in the end, the question asked by Gilbert Krulee in 1973 remains essentially unanswered. Everyone could agree that the attitudes of parents and teachers had some influence on children and classroom outcomes, and so obviously one ought to be interviewing parents as well as teachers; but what did one want to learn from these interviews? It could be argued that efforts to analyze parent and teacher data in national Follow Through evaluation studies foundered because of practical problems--the "volatility" or unreliability of parent responses, and the high rate of turnover of teachers in the program. Yet more fundamentally, these endeavors seem to have been undermined from the start by ambiguity of purpose. Occasionally outside observers pointed attention to the ambiguity of purpose behind the parent and teacher surveys. But since the focus of attention in the national evaluation since 1972 has been on estimating program effects in terms of children's test scores, these ambiguities were never brought to center stage and resolved.

---

[7] Judy Crooks and Donald Burnes, memorandum "Consolidated Reivew of September 17 Deliverables" to Mr. Fred Bresnick, Acting Chief Research and Evaluation Section, October 6, 1971.

IV. SUMMARY AND CONCLUSIONS

> The decision to use the very limited funds available for Follow Through...to initiate a program which will permit examination in depth of the consequence of different program approaches holds promise of inaugurating what could be literally a new era in governmental support for educational and social ventures, i.e., an era in which the knowledge and technical expertise of the educational specialist, the systems engineer, and the behavioral scientist are brought into harmony with the pluralistic value structure of our society.
>
> -John Hughes, Director
> Division of Compensatory Education
> Office of Education, 1968[1]

The national evaluation of Follow Through has come a long way since 1968 and the Pittsburgh evaluation. Since then the evaluation effort has drawn on the energies of literally thousands of people and has cost roughly $50 million. Scores of evaluation reports have been produced and together they run to literally tens of thousands of pages. The purpose of this technical history has been to review the national Follow Through evaluation studies, with particular emphasis given to nine major evaluation reports. The focus of the review has been the manner in which these evaluation studies have handled key analytical issues which troubled systematic evaluation of the program.

This summary of the analysis report and conclusions deriving from it are divided into six parts which describe: (1) the connection between the

---

[1] "Follow Through Budget and Organizational Staffing Requirements for FY 1968 ($15,000,000)," memorandum from John F. Hughes, U.S. Office of Education, to Notal Estes, associate commissioner for elementary and secondary education, Jan. 24, 1968. Quoted by Richard Elmore, "Design of the Follow Through Experiment," in A.M. Rivlin and P.M. Timpane (eds.), Planned Variation in Education: Should We Given Up or Try Harder? (Washington, D.C.: Brookings, 1975), pp. 23-45.

FT program and the national evaluation; (2) the scope and focus of the evaluation; (3) the way in which analytical issues troubling the evaluation of student outcome data have been handled; (4) the manner in which parent interview and teacher questionnaire data have been gathered and analyzed; (5) a review of substantive findings from the evaluation studies; and (6) comments on the accomplishments and failures of the national FT evaluation. The first four parts summarize the foregoing four chapters and the latter two recount more general concerns emerging from them.

The Follow Through Program and the Planned Variation Evaluation. The Follow Through program was born out of ambiguity and raised under conflicting expectations. At least in part it was motivated by the uncertain contention that a follow through on Project Head Start would help consolidate the benefits of compensatory education and so better promote the life chances of disadvantaged children. Yet the gains on which Follow Through was to build were never clearly identified.

Initially the program was caught in a sort of limbo between two federal bureaucracies and subsequently it has been torn between different conceptions of the program's main purpose. Follow Through was authorized under the 1967 amendments to the Economic Opportunity Act of 1964, and so became the official responsibility of the Office of Economic Opportunity. Even before the program began, however, operational responsibility was shifted to the Office of Education. Follow Through was conceived as a nationwide service delivery and community action program, but when Congress limited funding for such programs in late 1967, it was transformed--partially--into a planned variation experiment. The planned variations were to be alternative educational models developed and implemented by a variety of sponsoring agencies. Yet since the whole notion of sponsored planned variations was brand new, Follow Through also became a research and development effort for the sponsors to refine their models and improve methods of implementing them in FT projects around the country.

The changes in bureaucratic responsibility for FT and the various aims for the program led to controversy over the main meaning of FT which has yet to be resolved. Is Follow Through mainly a social action program, a service delivery program, an R & D effort to develop alternative approaches to compensatory education, or a systematic evaluation of alternative educational models? Writing in 1973 the original Follow Through director Robert Egbert lamented this ambiguity as a "fundamental and pervasive problem,"

> Administrative decisions have never been fully rationalized with the authorizing legislation--that is, the legislation authorized a social action program; the executive branch decreed that the program should be experimental. Legislators, community action agencies, OEO regional offices, and others were never apprised of this decree except by Follow Through personnel. Since community action agencies, regional OEO personnel, and local community people were accustomed to all kinds of rationalization for the failure of programs to meet community action standards, they were skeptical of this interpretation. Moreover, in spite of the decision that the program be experimental, the executive branch of the government often treated it as a social action rather than an experimental program.[2]

The FT evaluation was caught in this ambiguity and stretched between the different conceptions of the program. An early evaluation, in the pre-planned variation days of FT was conducted by investigators at the University of Pittsburgh. Their effort was so ambitious and hurried, however, that most of it fell through.

---

[2] Robert Egbert, "Planned Variation in Follow Through", p. 19. Egbert also mentions a number of other fundamental problems: "Failure of both policy-makers and program personnel to be explicit in their expectation of Follow Through;...priorities have changed, both within the communities being served and within the federal government;...Follow Through was in an inappropriate positon in the organizational structure." (pp.20-21.)

The first planned variation evaluation contract was awarded to the Stanford Research Institute of Menlo Park, California. From the very start the SRI evaluation was badly rushed; the evaluation contract was signed at the end of June and testing was to begin in four dozen sites across the country in fall 1968. As the evaluation was beginning that fall, both program and evaluation were attacked as insensitive to social action aspects of the program. So several new model sponsors representing minority institutions were added to the program and the evaluation was broadened to include case studies of institutional change as well as more traditional student testing.

In 1969, the evaluation came under fire again, as biased in favor of more traditional educational approaches. Responsibility for shaping the evaluation was somewhat ambiguously shared between OE officials and SRI investigators, and as the evaluation came under criticism new tasks were added to SRI's work--noncognitive instrument development, for example--to encompass the varied perspectives on the program. As the evaluation effort was stretched between the different conceptions of the program, ambiguity prevailed as to priorities. An outside review of the evaluation in 1970 criticized the efforts "diffuseness" and "mindless empiricism."[3] Uncertainty over the course of the evaluation heightened when early drafts of SRI reports were deemed unacceptable by OE officials, and when other external reviews of the evaluation effort proved highly critical.

All of this, plus changes in OE personnel responsible for FT, led to several important changes in the evaluation in late 1971 and early 1972. The evaluation was focused more narrowly on a comparison of the effects of the planned variation models in terms of a set of four standard tests. Also, although SRI continued to collect the FT evaluation data, the Office of Education sponsored a competition to find a new contractor to analyze the evaluation data. The analysis contract was awarded to Abt Associates of Cambridge, Massachusetts in June 1972. During the first phase of evaluation analysis at Abt, roughly 1972 to 1974, controversy continued over the main meaning of FT, but the evaluation was largely insulated from it. The Abt evaluators' job was clearly circumscribed to the analysis of the data which were largely faits accomplis. The data had been gathered by SRI according to the design formulated by OE and OE consultants at the Huron Institute, and employing standard test instruments selected by a special instrument selection panel.

---

[3] SRI, Draft Administrative History, pp. F-19.

In 1974, the contract for analyzing the FT evaluation data was again opened for competitive bidding and again was won by Abt. The second phase of evaluation work at Abt focused even more directly on planned variation comparisons. Deleted from the schedule of questions was the issue of whether the FT program in the aggregate has a greater impact on participants than non-FT schooling--the question which had headed the list of issues to be addressed during the first phase at Abt. During the second phase at Abt, controversy continued over the main meaning of FT--erupting regularly at annual conferences held by FT sponsors--but the analysts' work remained largely insulated from it. Their job was still circumscribed to analyzing data--from students' tests, parent interviews, and teacher questionnaires--which had been gathered with instruments and by a design which had been much disputed, but by then were beyond alteration. So the issues confronting the Abt analysts lay largely in the realm of data analysis: How could the available data be analyzed to yield valid estimates of the effects of the FT planned variation models?

A host of analytical issues had to be handled, but the one which achieved most prominence was the preschool confounding issue. FT was designed to benefit disadvantaged children who had previously attended Head Start. So more FT children than comparison children had attended some sort of preschool before entering kindergarten or first grade. The problem then was how to adjust for this difference in comparing third-grade test results for FT and non-FT children. In part because of worry over this issue, the announcement of results of a major Abt report in early 1976 was postponed--the summary chapter was simply deleted from the final version of the Abt III report.

So the elaboration and summing up of the national FT evaluation results was postponed to rest almost entirely with the Abt IV report. This _magnum opus_ of the FT evaluation was completed in April 1977, one full decade after President Lyndon Johnson had originally proposed a follow-through program during the first school grades for children in cases of acute poverty.[4]

Scope of the Planned Variation Evaluation. Since around 1970, the FT evaluation has undergone a sort of funnel vision, as dozens of questions which have been asked of the evaluation at one time or another have fallen by the wayside. Part of the cause for the cloud of different questions asked of the evaluation is that no single definition of the FT program itself has ever been universally accepted. Although FT was originally authorized in the Economic Opportunity Amendments of 1967, official rules and regulations for the program were not issued until April 1975. And even then the official definition of the program was not skyblue clear:

---

[4] Actually, Abt's analyses were not entirely brought to a close in April 1977. Four volumes of the Abt IV report were finished in April but Abt then received a special continuation contract to go back and investigate the preschool confounding issue more thoroughly on data from three cohort-streams of children which had previously been analysed. These follow-up analyses are to be reported in Abt IV volumes E and F, scheduled for completion in September 1977.

> The Follow Through Program...is an experimental community services program designed to assist, in a research setting, the overall development of children enrolled in kindergarten through third grade from low-income families, and to amplify the educational gains made by such children in Head Start and other similar quality preschool programs...[5]

Whatever FT has been exactly, it has been a big effort, involving roughly 20 model sponsors working in one or more of the 170 FT projects around the country. The FT sponsors range widely: from the Open Education approach of the Education Development Center, to the Behavior Analysis approach sponsored by the University of Kansas; from the mainly classroom-based Direct Instruction approach of the University of Oregon, to the Parent Education model of the University of Florida. Not all FT projects are affiliated with sponsors, however. Some 15-30 projects have existed over the years as either self-sponsored or parent-implemented programs.

Amidst this diversity the evaluation has ranged widely. In the early years, the tendency seemed to be to address a diverse set of evaluation questions in a few settings. Yet in the controversies over the evaluation and main meaning of the program, many early efforts eroded away with scarcely a trace remaining. Early attempts to develop special noncognitive test instruments, to conduct case studies of institutional change and to survey systematically FT Policy Advisory Committees, all were discontinued by 1971 without ever getting described in any final report.

Some evaluation questions lasted longer, withering in prominence over time, but with their skeletons remaining to be seen even in the final Abt report. For example, various efforts have been made to assess the implementation of sponsors' models in local FT projects--notably Jane Stalling's 1972-73 classroom observation study. But despite several further attempts to assess implementation, the goal of linking implementation evaluation with program effects estimations in the national evaluation was never accomplished. So in the final Abt report only the bones of the implementation issue remain as warning to future investigators.

---

[5] OE, "Follow Through Rules and Regulations," Federal Register, Vol. 40, No. 77 (April 2, 1975), part II, p. 17712.

> An unknown proportion of the variation [in apparent model effects in different sites] results from implementation problems, of which we have no analytically useful record...If future evaluations of externally-guided compensatory interventions wish to be able to attribute outcomes more confidently to the specific kinds of intervention applied, they will have to take steps to find out, systematically, what local people did with the models and other associated program resources. Model labels have turned out to carry only a small fraction of the relevant information.[6]

Other evaluation questions--concerning program effects on parents and teachers, for instance--gradually faded from prominence in similar fashion. As the national evaluation narrowed to focus almost entirely on a comparison of model effects on a set of standard outcome measures, the scope of the national evaluation did broaden somewhat in two ways. First, the later reports included data on a wider range of sites and sponsors than earlier reports. The Abt IV study, for example, reported on outcome data from 100 different sites--almost twice the number included in any previous report. Second, the later outcome evaluations (the Abt III and IV studies), reported data which spanned the full four years of the program--kindergarten to third grade--whereas earlier studies treated mainly one- or two-year spans of data.

So in terms of range of projects and time-span covered, the national evaluation broadened toward the end. This breadth of coverage was, however, accompanied by a narrowing of focus on alternative model effects on children's third-grade test scores (and in the RMC study the "costs" of alternative models). Omitted were the descriptive questions; (In terms of real practice what are the FT program and the different sponsor models?) and almost all issues of process (How do the program and the sponsor models work to produce whatever effects they do have?).

---

[6] Abt IV A, pp. 166-167.

The focus on effects questions stemmed from many factors--pressures on the evaluation to produce findings, the past difficulties in pursuing process questions, and concerns for showing the effectiveness of the program. Yet perhaps the main reason for this focus was the way in which evaluators tended to view program evaluation.

Analysis of Student Outcome Data. A primary cause of the funneled view of the national FT evaluation is the paradigm which evaluators have used to judge Follow Through. The paradigm, which is clearly predominant among program evaluators in general, is the treatment-control model of experimental research. The idea is simply that if we start out with two groups of experimental subjects which are essentially the same, apply a treatment to one group and withhold it from the second, then we can simply measure the differences between treatment and control groups at the end of the treatment period, and infer that the difference is an effect of the treatment.

The paradigm is seductively simple but has raised a bewildering array of problems when applied to the national FT program. First of all, FT and non-Follow Through comparison groups were not selected in the same manner so they tend not to be essentially equivalent at the start of kindergarten. The ideal of the treatment-control experimental model is that groups be randomly selected and assigned, but randomization was tried in only one small corner of the FT evaluation and even there did not prove much of a panacea. In general, selection and assignment have operated in FT in a variety of complex and not very well-documented ways. Even Follow Through sponsors were selected in several stages. Fourteen were chosen in 1968; six more were added in 1969-70; one dropped out and four more entered the program in 1970-72; and two dropped out in 1975. In similar fashion FT projects were added in several stages from 1968 to 1972 and several have dropped out since that time. Grounds for assigning sponsor models to projects likewise varied over time, although discretion for affiliations between sponsors and projects almost always seems to have lain with local representatives of the projects. Within projects, grounds for selecting schools and students to participate in the program also ranged widely.

The selection issue did not end there either, because not all FT projects or schools were included in the national evaluation. The evaluation sample was picked for a variety of reasons which, like all other stages of selection, seemed to vary from case to case and time to time. Finally, only after all of these stages of selection were accomplished were NFT comparison groups chosen. Like all else, the degree of similarity of evaluation sample projects and their comparison groups varies widely by case, but in general, the comparison sample tends to be somewhat richer and to have attended preschool in smaller proportions than the FT sample.

Given this hierarchy of entities in FT--sponsors, projects, schools, classes, and children--evaluators have had to decide at which level to conduct analyses. Early FT evaluators tended to use the classroom or the school as units-of-analysis in estimating program effects. Later analyses of FT evaluation data have tended to use the individual student as the unit-of-analysis. The grounds for using alternative units-of-analysis have been widely debated, both in the FT evaluation literature and elsewhere. Nevertheless, the switch from using the class to using the student as the unit in FT evaluations seems mainly due to one simple fact. Children's classes normally do not remain intact for more than a single school year, and the later FT evaluations spanned four school years.

An even bigger problem in applying the treatment-control paradigm in the FT evaluation has been how to determine what the FT "treatments" or models really are. FT model sponsors did not begin in 1968 with exact prescriptions of how children should be educated in the kindergarten through third grade years. Rather, they started with more general philosophies and practical predilections on how children should be educated. So FT evaluators tried both to work with sponsors to specify more precisely what the FT models are, and to verify that projects nominally affiliated with particular sponsors were in fact implementing the sponsor's model or approach. These efforts were troubled in several ways. Sponsor personnel sometimes changed, and some of their approaches shifted over time. Some sponsors, especially the ones with more open approaches, also simply reject the notion that their models can effectively be captured in any explicitly prescribed operational terms.

So the fear of many has been that the evaluation may be dealing in illusions, assessing nominal treatments which do not exist in practice. The problem has been discussed in all of the outcome evaluations--those carried out by SRI investigators, and by Abt analysts in both the first and second phases at Abt--but all to little avail. In their final report Abt investigators noted simply that "There has been no adequate assessment of whether the sponsors' evolving programs were implemented."[7]

The implementation issue has been addressed more directly in FT classroom observation studies. Studies by both Jane Stallings and Robert Soar have shown that observed classrooms affiliated with different sponsors do tend to differ in ways that correspond to sponsors' stated approaches.

---

[7] Abt IV A, p. 10.

Yet since there have never been precise descriptions of exactly what fully implemented classrooms of the different models should look like, even the classroom observation studies have been able to make little progress on the implementation issue.

The other side of the treatment-control paradigm is what happened to the control groups with whom treatment groups are compared. Obviously, NFT comparison groups are receiving some sort of schooling, but there has been no way to tell exactly what it is that the FT models are being compared against. Not only have NFT comparison children been receiving some kind of schooling, but most have been attending schools which receive federal funds for compensatory services under Title I of the Elementary and Secondary Education Act of 1965. In addition, anecdotal reports suggest that in at least some cases, aspects of sponsors models may have diffused to be adopted in the NFT schools. Moreover, the classroom observation analyses reported by Jane Stallings showed that NFT classes appear to vary nearly as widely as FT sponsors' classes.

Linked to the problem of defining FT treatments is the issue of how long the FT treatments and the evaluation of them should last. Since FT was originally designed to build on the benefits of Head Start, it could be argued that the "treatment" of interest is not just FT, but the combined experiences of HS and FT. However, since the FT evaluation was designed almost entirely separately from HS evaluations, evaluators have been forced to treat entry into FT as the start of the treatment. At the other end of the program, people have advocated both that the FT program should be extended beyond the third grade and that the evaluation, even if not the program, should follow children beyond the third grade. The concern has been that the long-term goal of FT--enhancing children's life-chances--is simply not adequately measured by children's third grade test scores. People have worried that apparent gains at the end of third grade may prove illusory and fade out after children leave the program, and that deeper, more complex, effects of FT may not show up until later in children's lives. In any case, the FT evaluation has been limited to assessing program effects which can be measured in the spring of children's third grade school experience.

What program effects can and should be measured at the end of third grade has raised even more dispute than the temporal issue. In the early years, FT evaluators attempted to develop new test instruments with which to assess the diverse goals of the program and the varied aims of the different model sponsors. However, when this endeavor proved largely fruitless, and the evaluation effort fell under criticism, a standard set of test instruments was chosen to use in testing third grade FT students. The set included the Metropolitan Achievement Test (MAT) the Ravens Coloured Progressive Matrices, and two noncognitive instruments: the Intellectual Achievement Responsibility Scale (IARS) and the Coopersmith Self-Esteem Inventory.

Though this standard test battery was chosen in 1972, considerable effort has since gone into developing alternative test instruments for FT. In the national evaluation two separate efforts were made to find subsets of items from the MAT which might show particular effects of FT. Also, FT sponsors have been working to develop a variety of instruments for assessing their programs in terms broader than group-administered written standardized tests.

Another issue troubling the evaluation has been the attrition of children out of the program--or at least out of the sample of tested children. The problem is that not all of the children tested at entry into the program have been tested at the end of third grade. The problem did not seem terribly severe in early evaluations spanning only a single school year or in ones using the class as the unit-of-analysis, but in recent evaluations over the full kindergarten to third grade span of the program, it has proved more troublesome. Only 30-35% of the children tested at entry into kindergarten were still in the program and successfully tested in the spring of their third grade. The issue of how this may bias evaluation results has been treated differently in different evaluation reports, but it seems clear that in at least some sites with differential attribition out of FT and NFT groups, the problem may seriously affect program effects estimates.

The primary analytical tool for evaluating the effects of FT in terms of the standard test instruments has been analysis of covariance (ANCOVA). The technique has been used to try to adjust for initial differences between FT and comparison groups of children. The ANCOVA has variously been applied to estimate overall FT effects, sponsor effects, and site-level effects. Schools, classes and children have alternatively been treated as subjects or units in these ANCOVA's. Different ANCOVA's of the FT data have also employed from three to twenty variables as covariates to adjust for initial differences between FT and NFT groups. Moreover, evaluators have examined the assumptions underlying the ANCOVA in widely different fashion. All of the variations in the application of ANCOVA suggest that the results of no one analysis ought to be too much trusted. Rather, it seems as if meaningful evaluation results can be derived only from a balanced consideration of a variety of them.

Analyses of Parent Interview and Teacher Questionnaire Data. Although the FT evaluation has focused mainly on children's test data, several million dollars have been invested since 1968 in collecting data from FT parents, teachers and teacher aides. These efforts have been especially troubled by lack of continuity in data-gathering instruments, ambiguity of purpose and variation in how the data were analyzed.

As part of the Economic Opportunity legislation, Follow Through was mandated to have parents involved in several aspects of the program. In the national evaluation the main mechanism for assessing their involvement has been interviews carried out in person with FT parents. The forms used for parent interviewing were repeatedly changed from 1968 to 1972, but were not altered after that. In the spring of 1973, 1974, and 1975, 87-question interview forms were administered to several thousand parents of third grade FT and NFT children. Though the format stabilized in 1973, the issue of exactly why the parent data were gathered and how they would be used have not proved so easy to pin down. Though accounts of the purpose of parent-interviewing differ, basically three different types of reasons have been given.

> The parent interview data would provide background data on FT and NFT parents and children to be used in analyses of children's test data--both as adjustment covariables in estimating treatment effects and to explore whether different types of approaches seem differentially effective with different types of children.
>
> The interview data would be used to assess the impact of the program on parents themselves.
>
> The data would provide a means of assessing the degree to which the program approaches had actually been implemented.

The only purpose which seems to have proven out with any consistency is the first. Almost all of the analyses of children's test scores have used one or more items of data from the parent interviews as covariables to adjust for initial differences between treatment and comparison groups.

Using parent interview data to assess program impact on parents seems to have failed for several reasons. Although the parent-interviewing was organized in pre-post fashion, the entry and exit interview forms administered to parents were quite different. Moreover, only one attempt to link the pre- and post-interview data has been made, and that did not prove very successful. Also, the methods used to analyze the parent data in the different major evaluation reports (mainly factor and cluster analyses) have yielded different sets of synthesized variables, which often are not the same from one report to the next.

The use of parent interview data to assess program implementation seems to have fallen by the wayside for several reasons. First, exactly how the parent data would be used to assess which aspects of program implementation was never clearly spelled out. Second, the validity of interview data in terms of what actually happened in FT projects was never critically examined. Third, the parent data were analyzed in such a variety of ways and yielded such abstract variables as to make their interpretation in terms of program components virtually impossible. As a result, and as the FT evaluation neared completion, the parent interview data were linguistically devalued. Originally the parent data were to provide empirical estimates of implementation, but in the final Abt report were termed merely indications of parents' acceptance of model tenets.

Although teachers and teacher aides obviously play a crucial role in implementing educational programs, they have received relatively little attention in the FT evaluation. Apart from the classroom observation studies, the only available source of information on them has been the teacher and teacher aide questionnaires. As with the parent interview data, use of information from these questionnaires has been troubled by variations in the information gathered at different times, ambiguity of purpose and inconsistencies in the way the teacher data were analyzed. In several respects, however, the history of the teacher and aide questionnaire data has differed from that of the parent data. First of all, it was gathered not in face-to-face interviews but by mailed questionnaires. Second, the exact purposes behind gathering the teacher data and plans for how it would be analyzed have been even less clear than those for the parent data. The teacher data has been linked to children's test data only in early evaluations using the classroom as the unit-of-analysis and analyzing data from only a single school year. Third, the teacher data gathered at different times has never been linked, so that it has been impossible to examine in detail how teachers' responses to questionnaire items may have changed over time.

The overall outcome of analyses of the teacher and aide questionnaire data has, however, been much the same as that of the parent data. Both have been used in reports outside the mainstream of the national evaluation, but in recent major national evaluation studies both parent and teacher data met a peculiar fate. A variety of abstract analyses (factor and cluster analyses) have been performed on the parent and teacher data. Considerable space has been given to reporting synthesized variables (factor and cluster scores) derived from them. Yet the meaning of the results of the analyses of the parent and teacher data remains quite murky. So the information on Follow Through parents and teachers faded into the text and appendices of the final Abt IV report and were not even mentioned in the summary chapter or in the digest of this magnum opus of the Follow Through evaluation reports.

Overview on Findings. Given all of its problems, what has been learned in the national Follow Through evaluation? What are the findings from a decade-long evaluation effort that cost $50 million? Synthesis of findings has not been a primary goal of this report, but after 300 pages of analyzing the problems of the national Follow Through evaluation, some summary of what seems to have been learned is clearly appropriate.

First of all, it should by now be obvious that there is no easy or mechanical way to sum up the findings from the various national evaluation reports. Different analyses have been based on different samples, have encompassed different spans of the program, and have used widely varied analytical techniques. It is impossible even to sum up the summaries of the major evaluation reports because the Abt I and Abt III reports simply had no summaries. Therefore, the approach employed here is not highly rigorous. Rather, it is more like fitting available pieces of a puzzle together to show the outlines of a picture. Instead of yielding definitive findings, it can give us only probable conclusions.

The obvious place to begin review of evaluation findings in this realm is the Abt IV report. The "findings" from this report are listed in Table 14. Included in this table are the ten findings reported in the Abt IV A volume and the four summary findings from the volume IV D longitudinal study. Note that the last item of the IV D summary is really a recommendation rather than a finding.

The major ambiguity underlying all of these findings is the rather basic issue of how to infer whether there are any significant model effects on any outcome measures. The strategy followed in both the Abt IV A summary of results and the IV D longitudinal study is simply to aggregate

Table 14: Findings of the Abt Volume IV Evaluation Report

Ten Findings Reported in Abt IV A

1. The effectiveness of each Follow Through model varied substantially from site group to site group; overall model averages varied little in comparison.

2. Models that emphasize basic skills succeeded better than other models in helping children gain these skills.

3. Where models have put their primary emphasis elsewhere than on the basic skills, the children they served have tended to score lower on tests of these skills than they would have done without Follow Through.

4. No type of model was notably more successful than the others in raising scores on cognitive conceptual skills.

5. Models that emphasize basic skills produced better results on tests of self-concept than did other models.

6. Model comparisons in New York and Philadelphia yield results which are similar to those found in overall comparisons.

7. Some models are more successful in their most disadvantaged sites.

8. Two models are consistently more effective with Head Start children.

9. Most models are more effective during kindergarten and first grade than during second and third grade.

10. Some Follow Through sponsors grew in effectiveness over time.

Four Findings Reported in the Abt IV D Longitudinal Study

11. Follow Through programs have been relatively more effective in the quantitative area than in the verbal area on group academic achievement tests, particularly during the first two years of elementary school.

12. A variety of educational models are capable of producing notable progress in the early period in the quantitative areas. In the verbal area, most positive effects on academic achievement tests derive from programs with a direct academic emphasis, and where these effects persist, they do so longer in this type of program.

13. The tendency to gain relatively less over the first two years appears more frequently for Head Start groups than other preschool groups.

14. Proliferation of innovative educational models should be accompanied by corresponding non-traditional measures and research designs aimed at capturing more fully the range of educationally active ingredients.

Sources: Stebbins, Abt IV A, pp. 135-155, and Ferb, Abt IV D Longitudinal Study, pp. 91-92.

effects estimated at a lower level (mainly site-level effects in IV A, and sponsor preschool-group effects in the IV D study) and to refer to the aggregations, either directly or indirectly as model effects. The two studies do this in slightly different ways. The IV A summary averages trusted positive, null, and negative effects together to estimate model effects, while the IV D summary simply considers the proportion of positive or the proportion of non-negative trusted effects for each model. This is not a major difficulty, however, since the different methods of aggregation employed in the IV A and IV D summaries can be reconciled from the information given.

The really major difficulty in either case is how model effects can be inferred in light of the apparent variation in effects estimated at the disaggregated levels. A devil's advocate could, in fact, argue that the national evaluation tacitly admitted that there were no model effects after the authors of the Abt II report gave the site variability issue special stress in their final recommendation for a series of "detailed studies at the site level,"[8] and analysts in the second phase at Abt stopped estimating effects at the sponsor level, and started estimating effects only at the site (or sponsor preschool-group) level. What the Abt IV summary implicitly does is say that a model "effect" is simply the average of site-level effects estimates for all sites affiliated with that model. The more traditional statistical approach would be to ask whether the magnitude of apparent model effects is large enough in light of the overall variation in estimated site-level effects to warrant the inference of any significant model effects. For example, one might perform a one-way analysis of variance (ANOVA) on the site-level estimated effects to see whether affiliation with sponsor models accounted for a significant proportion of variance in the site-level effects. In light of Abt's Figure 5-2 summary of results, it seems almost sure that such an ANOVA would reveal that affiliation with sponsor (the main effects of the ANOVA) overall explains an insignificant proportion of the variation in site-level effects. Such a test would, however, be a very tough test of model effects. There are so few sites affiliated with some sponsor models that such a test would have little power to reveal sponsor effects even if they were real.

---

[8] Abt IIA, p. VII-10.

Variability among Follow Through sites has been noted in studies other than outcome evaluations, too. For example, in his review of health care delivery in Project Follow Through, Thomas Sullivan noted that "Because of the variances in local circumstances from community to community, the achievements in one community would appear as nothing in another."[9] So the essence of Abt IV A Finding #1 seems quite strongly supported by a variety of other evidence: effects estimates, and a lot of other things as well, vary widely among Follow Through sites, even those affiliated with the same sponsor.

The second finding of the Abt IV study is "models that emphasize basic skills succeeded better than other models in helping children gain these skills." By "models that emphasize basic skills," the authors of this finding refer to the Direct Instruction, Behavior Analysis, SEDL Language Development, and the California Process Models. "Helping children gain these skills" is assessed strictly in terms of children's scores on four subtests of the MAT: the word knowledge, spelling, language and math computations subtests. The Abt analysts do not directly discuss the question of how much basic skills models appear to boost children's basic skills test scores, and whether the magnitude of apparent sponsor effects on basic skills tests seems significant in light of site-to-site variation on these tests. What they do note is the following:

---

[9] Thomas T. Sullivan, "Health Care Delivery in Project Follow Through," (Cambridge, MA: Bio-dynamics, Inc., November 16, 1970), p. 1.

> It is reasonably clear from these results that if an educator is aiming primarily at enhancing spelling, vocabulary, and arithmetic computation skills, one of the basic skills models would be a better bet than the others. Our earlier finding of wide site variability still holds, however: even the model with the best track record had more than one site in which the Follow Through children did less well, on the average, than we estimate they would have done if they had been outside Follow Through. Unmeasured local circumstances, including those associated with implementation, still have had more influence on results than have the philosophies of the models' sponsors.[1]

Despite this warning, the finding that models stressing basic skills seem to help raise children's basic skills test scores appears to be stronger than any of the remaining findings. Even so, there has been little consistency over the years in which Follow Through models have been classified and evaluated as basic skills approaches. For example, the Abt IVA volume classifies the SEDL Language Development and California Process models as basic skills approaches. In contrast, the authors of the Abt IVD longitudinal study declined to include these two models in a comparative evaluation of outcomes because of their special characteristics:

---

[1] Abt IVA, p. 146.

> The comparison of SEDL to the other FT programs is inappropriate since this program is unique among the FT programs in that it was designed to serve bilingual and/or Spanish-surnamed children. The effects that the SEDL program has had would not be expected to occur with other groups, e.g., other types of minority children, nor would other programs not designed to deal with bilingual children be expected to produce effects similar to those of the SEDL program. Ideally, the progress of the SEDL children should be compared to the progress of similar children involved in other programs designed to serve bilingual children or to the progress of similar children in traditional classrooms.
>
> The CAL Process group, as noted, is quite small and also atypical with respect to background characteristics....[T]hat the children included for analysis are substantially different from other FT models' groups is sufficient reason to exclude the sponsor from comparative study.[2]

Therefore, it is important in summarizing Follow Through findings to specify clearly which Follow Through models are included in which conclusions. If evaluation findings are considered on a sponsor-by-sponsor basis, it seems clear that the Direct Instruction model of the University of Oregon shows up as more effective than other Follow Through models in terms of boosting children's third grade basic skills test scores. As the text of the Abt IVA summary chapter notes, "the Direct Instruction Model of the University of Oregon...had a distinctly higher average effect on basic skills scores than did any other model...."[3]

---

[2] Abt IVD, p. 72. These comments were made concerning the sample treated in the longitudinal study but are pertinent also to the broader cohort II-EK sample treated in volume IVA. Compare IVA pp. 51-52 with IVD pp. 17-19.

[3] Abt IVA, p. 146.

Since the Direct Instruction model appears to have won the evaluation horse race in terms of children's third grade test scores, it is useful to look more closely at the dimensions on which this model appears to have had an impact. In the Abt IV summary of results, four MAT subtests were classified as basic skills measures:

- Word Knowledge Subtest (50 items)
- Spelling Subtest (40 items)
- Language Subtest (Parts A&B together 50 items)
- Mathematics Computations Subtest (40 items)

Examination of Abt's "Summary Effects For Direct Instruction Sites" shows that effects estimates for this model were not uniform across these four "basic skills" measures. Rather, effects in this domain seemed concentrated on the Language and Mathematics Computations subtests.[4] This pattern was also evidenced in an appendix table which reports for each model the percentage of variance explained in each outcome measure by demographic covariates, between site covariates and sponsor "treatment" terms for the cohort III-K local analysis.[5] This table indicates that in the local analysis the Direct Instruction "treatment" term accounted for more variance in children's language and math computations subtest scores (10% and 9% respectively) than any other sponsors on any other outcome variables. The table also indicates, however, that when the two parts of the MAT language subtest are viewed separately, quite different results emerge. In the cohort III-EK local analysis, the Direct Instruction treatment explains twelve percent of the variance in children's language part B scores, but only one percent of the variance in their language part A scores.[6] Part B of the MAT language subtest deals with punctuation, capitalization and usage while part A tests children's ability to interpret the sense of sentences. So in sum these considerations suggest that the impact of the basic skills Direct Instruction model is more apparent on the tests of the most concrete of the basic skills (math computations, punctuation, capitalization and usage) than on tests of more complex skills.

---

[4] Abt IVB, p. 72

[5] Abt IVA, pp. 190-93.

Despite these specific considerations, the more general finding that basic skills models tend to boost children's basic skills test scores finds support in other Follow Through evaluation studies. For example, from his classroom observation study and examination of the relationships between classroom behavior and pupil test score gains, Robert Soar concluded,

> Probably the most consistent finding was a relation between a number of factors from the five [classroom observation] instruments indicating that greater amounts of teacher control, structure, focus, and convergence, or lesser amounts of pupil freedom, exploration of ideas, or experimental teaching led to increased pupil cognitive growth, especially in the skill measures.[7]

Jane Stallings in both her 1971-72 and 1972-73 studies reached similar conclusions. In her earlier study she reported,

---

[6] The difference in apparent Direct Instruction effects on the two parts of the MAT language subtest can be seen also by comparing pp. 272 and 273 of Abt IVC.

[7] Soar, Classroom Process, p. 203.

> Results indicated that the children of...sponsors who emphasize behavior modification processes [the Oregon, Kansas and Pittsburgh sponsors] perform better on tests related to traditional academic skills than children of other sponsors.[8]

From her second and larger study she reached a similar conclusion,

> The effect of classroom processes upon children was examined through partial correlations and stepwise regressions. Since the classroom processes predicted the outcome scores as well as or better than did the entering baseline test scores, we conclude that what teachers do does make a difference. In the more academically oriented classrooms which use a high rate of drill, practice, and praise and have the children more frequently engaged in reading or math activities, the residual gain scores on reading and math are higher.[9]

The finding that basic skills models help boost children's scores on basic skills tests contrasts with Abt's finding 4: "No type of model was notably more successful than the others in raising scores on cognitive conceptual skills." By "scores on cognitive conceptual skills" the authors of the Abt IV A study refer to the Ravens, and the reading, math concepts and math problem-solving subtests of the MAT. The contrast in findings for these outcomes with those for the basic skills outcomes is also revealed in the fact that the sponsor treatment terms in the cohort III-EK local analysis accounted for no more than 4% of the variance in any of the third grade cognitive skills test scores.[1]

---

[8] Stallings, Classroom Observation 1971-72, p. 265.

[9] Stallings, Classroom Observation 1972-73, p. 345.

[1] Abt IVA, pp. 190-93.

The distinction between simpler basic skills tests and more complex cognitive skills tests may also help explain the Abt IV D finding that "Follow Through programs have been relatively more effective in the quantitative area than in the verbal area on group academic achievement tests...." In the quantitative area, the Abt IV D longitudinal study included the MAT math computations subtest on which basic skills sponsors apparently had an impact, but did not include the language subtest, the measure in the verbal area on which basic skills models seemed to have greatest impact.[2]

Finding 9 that "most models are more effective during kindergarten and first grade than during second and third grade," also may be partially explained in terms of simpler versus more complex test content. The Abt IV longitudinal study analyzed changes in pupils' test scores between fall kindergarten and spring first grade (FK-S1) and between spring first grade and spring third grade (S1-S3). The test administered at spring first grade was the MAT Primary I while that given at spring third grade was the MAT Elementary. The apparent diminution in "effectiveness" of FT models in the latter period may be due simply to the shift in test content. The MAT Primary I contains a higher proportion of simple skills items than does the MAT Elementary test.[3] So Finding 9 may be a case in which the measure is the message.

---

[2] In the verbal area, the longitudinal study included only the MAT total reading score which was simply the sum of the reading and word knowledge subtests. Abt IV D, pp. 27-28.

[3] See Table 9, p. 188 for a description of the content of the MAT Primary I and Elementary tests.

There is another possible explanation of the apparent diminution of model effectiveness from the FK-S1 period to the S1-S3 period. The fall to spring school year lasts roughly eight months and summer vacation lasts, let us say, four months. Now if FT children are receiving the model "treatment" only during the school year, it is clear that the FK-S1 period encompasses an 80% strength treatment (that is children are in school sixteen out of twenty months in the FK-S1 period) while the S1 to S3 period covers only a 67% strength treatment (children are in school 16 out of 24 months). This sort of reasoning may also help to explain why the Abt I report found that the aggregate effect of FT across all sponsors for cohort III-EK children was both positive and statistically significant on almost all outcome measures.[4] In the fall to spring kindergarten period covered in the Abt I report, children were receiving 100% strength treatment--that is a straight eight months of schooling. In fact, in their kindergarten year FT children may have been receiving "extra-strength" schooling because many of them apparently attended FT kindergarten more hours than the half-day kindergarten which Non-Follow Through children attended.

No comprehensive information is available on how many hours a day FT children attended kindergarten, but anecdotal reports suggest that often it was more than the time NFT comparison children attended kindergarten. For example, in testimony before Congress, a Follow Through coordinator for the State of Washington recounted:

---

[4] Abt I, pp. VII-9-11. The Abt I report concluded that the one-year kindergarten "main effects" of Follow Through averaged across ten sponsors, on all eight outcome measures used at the end of kindergarten, were significantly positive and 0.1 to 0.4 standard deviations in magnitude.

> One of the purposes of Follow Through was to reinforce and sustain Head Start benefits as children move through the elementary grades. To have a child in the Head Start program for 3 1/2-hours a day this year and then move to a 2 1/2-hour regular kindergarten program in comparison with the 4 1/2-hour Follow Through program, with a loss of services is surely to have many of the gains achieved in Head Start this year disappear under such conditions. Head Start children need the continued enrichment and services offered by Follow Through.[5]

There is not sufficient information available to test whether differential amounts of schooling might account for the apparent success of Follow Through in the kindergarten year, but the hypothesis seems plausible and is consistent with other research on effects of schooling.[6]

Finding 5 of the Abt IVA report is one of the more intriguing ones of the entire Follow Through evaluation. It is that "models that emphasize basic skills produced better results on tests of self-concept than did other models." Tests of self-concept were the Coopersmith and the IARS positive and IARS negative. This finding is of special interest because it is consistent with a theory espoused by some of the basic skills sponsors; namely that enhanced self-esteem is the product of successful skill attainment rather than vice versa.[7] The finding is, however, somewhat dubious. Apparent model effects in this realm seem slight compared with site-to-site variation in effects estimates.[8] In the cohort III-EK local analysis, model "treatment" accounts for no more than three percent of the variance on any of the three self-concept measures.[9] Also, effects on the three different self-concept measures seem somewhat inconsistent.[1]

---

[5] Sharon Hunt, Statewide Action on Follow Through, Tacoma, Washington, testimony in U.S. Congress, House of Representatives, *Hearings on Extension of Certain Programs under the Economic Opportunity Act of 1964*. 93rd Congress, 2nd Session, Feb. 21, 1974, p. 108.

[6] For example, David E. Wiley, "Another Hour, Another Day: Quality of Schooling, a Potent Path for Policy" in Marcia Guttentag (ed.), *Evaluation Studies Review Annual* (Beverly Hills, CA: Sage, 1977, forthcoming). Before waxing over enthusiastic about this hypothesis, it should be noted that there is evidence among the Follow Through evaluation studies which seems partially inconsistent with it. In his *Classroom Process* report Robert Soar concluded: "One of the strongest relations in the study suggested that a long school day for entering first graders is negatively associated with gain" (p. 203). This suggests that if there is a relationship between amount of schooling and achievement, it holds only up to a certain limit; say maybe up to 4-5 hours of school per day for entering-level children.

Some of the larger estimated effects on self-concept measures show up on the IARS negative subtest. In several sites affiliated with the Direct Instruction and Behavior Analysis models, children's scores on the negative subtest of the IARS apparently were raised. In these sites children were, for example, more likely to attribute not doing well on a school test to their not doing their work, rather than to the difficulty of the test.[2]

---

[7] For example, Wesley Becker and Siegfried Engelmann, sponsors of the Direct Instruction model, wrote in 1971 that "Those who emphasize self-esteem as a primary goal in child development probably have the cart before the horse. Viewing self-esteem as the *goal* misses the point that self-esteem is the *product* of good teaching. One cannot teach attitudes about learning without presenting instances of learning tasks. One cannot teach self-esteem without presenting instances of accomplishment." Wesley C. Becker, Siegfried Engelmann and Don R. Thomas, *Teaching: A Course in Applied Psychology* (Chicago: Science Research Associates, 1971), p. 192.

[8] See Figure 5-2, Abt IVA, p. 136.

[9] Abt IVA, pp. 190-93.

[1] For the estimated effects for the four basic skills sponsors in the cohort III-EK local analysis and preschool adjustment analyses see Abt IVC, pp. 267-69, 361-63, 703-06, 821-24, and 282, 376, 718.

[2] This is an example of one of the questions in the IARS negative subtest.

This finding is not terribly clear in the Abt IV study, because as previously mentioned, there is a great deal of variation in site-level estimated effects on the IARS. Nevertheless, the hypothesis that models may have differential impact on three different affective measures (that is, the Coopersmith, the IARS positive and the IARS negative) finds support in Jane Stallings' 1972-73 classroom observation study. Stallings found very few significant relationships between instructional process variables and Coopersmith scores. She did, however, discover relations between IARS scores and the instructional processes, but they seemed quite different for the IARS positive and negative subtests.

> The Intellectual Achievement Responsibility [Positive] Scale shows a positive correlation with variables describing the more open classrooms. Our results indicate that children from the more flexible classrooms take responsibility for their own success, but not for their failure. Children in classrooms using the more flexible models of University of Arizona and EDC had higher adjusted scores than children in classrooms of other sponsors. Children from the more highly structured classrooms take responsibility for their own failure, but attribute their success to their teacher's competence or other forces outside themselves. University of Kansas and University of Oregon sponsored the more structured models, and children in classrooms using those models have higher adjusted scores than children in other sponsors. EDC was the only sponsor to have positive residual gain scores on both scales.[3]

These interpretations need to be viewed with considerable caution, however, because of the doubtful validity of general self concept measures. A recent review of self-concept measures, including the Coopersmith Self-Esteem Inventory concluded,

---

[3] Stallings, Classroom Observation, 1972-73, p. 344. See also Soar, Classroom Process, pp. 198-201.

> Considered as a body of research, self-concept studies lack the focus that would result from an agreed-upon definition of self-concept, lack adequate validation of interpretations of self-concept measures, and lack empirical data on the equivalence of the many self-concept measures currently being used.[4]

And in fact, the Follow Through data themselves and the intercorrelations of the various measures employed in the Follow Through evaluation cast considerable doubt on the validity of the self-concept measures used in the Follow Through evaluation. Correlational methods of test validation suggest that different measures of the same construct should correlate more highly than different measures of different constructs. If the Coopersmith and the IARS are valid measures of some general self-concept construct then they ought to correlate more highly than either would correlate with, say, academic achievement measures. Appendix Table A-4-1 of the Abt IVA report presents some direct evidence on this point. This table shows the intercorrelations of some two dozen variables used in the Follow Through evaluation for each of the three cohort streams of Follow Through and Non-Follow Through children.[5] In each of these six sets of data, Coopersmith scores correlate more highly with one or more of the achievement test scores than with either subtest of the IARS. In fact, in every case spring third grade Coopersmith scores correlate more highly with fall kindergarten WRAT scores than they do with third grade IARS-negative scores. This suggests that the supposed self-concept scores have more in common with academic achievement tests than they do with each other. In other words, the Coopersmith and the IARS may be measuring something more like paper and pencil test-taking skills than like any generalized self-concept. Whatever the case, the relatively strong correlations between the three affective measures and the achievement measures together with the relatively weak intercorrelations among the three affective measures themselves, suggests that results of analyses on the three non-cognitive measures should not be lumped together in an effort to generalize findings to any underlying unitary self-concept construct.

---

[4] R.J. Shavelson, J.J. Hubner and G.C. Stanton, "Self-Concept: Validation of Construct Interpretations," Review of Educational Research 46:3 (1976, 407-41).

[5] Abt IVA, pp. 178-183. Correlation matrices are presented separately for the following groups of tested children: III-EK FT, III-EK NFT, III-EF FT, III-EK NFT, III-EF FT and II-EF NFT.

Finding #6 of the Abt IV report, that "Model comparisons in New York and Philadelphia yield results which are similar to those found in overall comparisons," needs some elaboration to be understood. Since eight different models were operating in New York and Philadelphia, the Follow Through data from these two big cities offered a unique opportunity to compare model effects estimates in settings presumably more similar than in cases where different models were implemented in different cities. In other words, it was hoped that in these big cities the ceteris paribus assumption would more nearly approximate reality than in Follow Through comparisons overall.[6] For most models, effects estimates from New York and Philadelphia were lower than in overall nationwide comparisons--a finding consistent with sponsors' reports that it is harder to implement their models in these big cities

---

[6] Inspection of Abt's socioeconomic and ethnic-linguistic "discrepancy indexes" suggests that the variability between Follow Through and Non-Follow Through comparison groups may not actually have been much less in the Philadelphia and New York sites than in sites nationwide--at least in terms of background variables measured. See Abt IVB, pp. 4, 11, 12 and Abt IVC, pp. 23-8.

(and in their big school bureaucracies) than in smaller communities. Nevertheless, comparisons between models in these big-city sites were essentially the same as overall comparisons. Specifically, effects estimates on basic skills measures tended to be higher for basic skills models, particularly for the Direct Instruction model, than for other models.

The remaining findings of the Abt IV report can be mentioned briefly. In the text of volume IV the Abt authors suggested that findings 6-10 hold with "less certainty and generality" than their first five findings. This caveat seems especially warranted for findings 7, 8, and 10, for two reasons. First of all, these findings are of the glass-half-full-half-empty variety. From the evidence available one could almost as easily have said:

- Some models are less successful in their most disadvantaged sites;
- Some models are consistently less effective with Head Start children; and
- Some sponsors declined in effectiveness over time.

A more important limitation of these findings is that Abt's analyses were not much focused on these issues. Before one could have confidence in these findings he or she would need to consider carefully alternative hypotheses to explain apparent changes which at first glance have been interpreted in terms of differential effectiveness in disadvantaged versus relatively advantaged sites; with Head Start versus non-Head Start children; and with cohort II versus cohort III children. For example, it would be necessary to consider whether differential attrition or differential comparability between Follow Through and Non-Follow Through groups might be related to differential effects estimates on these dimensions. Hence, rather than placing overmuch faith in these auxiliary findings (specifically, findings

7, 8, and 10), it is probably more appropriate to view them in terms of supporting--or at least not being inconsistent with--finding #2.

* * * * * * *

One major national evaluation report has gone unmentioned since chapter II, namely A Cost Analysis of Follow Through Projects by Donald Rogers, et al., of the RMC Research Corporation. This report was not mentioned in chapters II or IV because it was carried out entirely separately from the analyses of children's test data and the parent and teacher survey data, and because it employed techniques which were almost entirely different from those used in other major evaluation studies. So before discussing the findings of the MRC report, let me digress briefly to provide some background.

The RMC cost analysis of Follow Through was an artifact of the enthusiasm in the 1960's for rationalized government decisionmaking. As manifested in Planning Programming and Budgeting Systems (PPBS), first in the Department of Defense, and later in other agencies of the federal government, the idea was that program decisionmaking should be based on empirical evidence regarding the effectiveness and costs of program alternatives.[7] Early plans for evaluating Follow Through reflected the tenor of the times. The Pittsburgh proposal for evaluating the pilot year of Follow Through, for example, held out as the ideal goal of that evaluation effort "definitive statements with respect to the relationship among input (programs and cost) and outcome (product)."[8] Evaluation of Follow Through costs was completely omitted from the pilot year evaluation, however, because, as the Pittsburgh final report explained:

---

[7] For fuller accounts of the relationship between the Follow Through experiment and the earlier enthusiasm for rationalized program management, see chapter 2 of Elmore, Follow Through: Decisionmaking or Alice M. Rivlin, Systematic Thinking for Social Action (Washington, D.C.: Brookings, 1971).

> [T]he cost evaluation was subject to the U.S. Office of Education being able to collect and verify such information. This collection was not done and accordingly no cost information is included.[9]

In the early years of the Follow Through planned variation evaluation at SRI, the issue of model costs seems to have drawn little attention. The SRI history of the evaluation notes that in a planning meeting on the evaluation in April 1970, one consultant "raised issues from the perspective of a school administrator regarding the analyses of program cost."[1] Initially, SRI investigators did not themselves attempt to gather cost data on Follow Through projects. Instead they relied upon cost information from the Elementary and Secondary Education General Information Survey (ELSEGIS), from the Consolidated Program Information Report (CPIR) and from Follow Through project application grants to derive estimates of per pupil costs.[2] In their Interim Evaluation report on 1969-71 data, Emrick and his colleagues calculated Follow Through expenditures per pupil for each of 39 projects on which data were available. They found that expenditures ranged from

---

[8] Parenthetical remarks in original. Quoted in Kurfeerst and Stephens, p. 1.

[9] Kurfeerst and Stephens, p. 1.

[1] SRI, Draft Administrative History, p. F-21.

[2] ELSEGIS and CPIR were parts of the ill-fated Belmont Joint (Federal/State) Coordinated Evaluation System. For a description of Belmont and its demise see Milbrey W. McLaughlin, Evaluation and Reform: The Elementary and Secondary Education Act of 1965, Title I (Cambridge, MA: Ballinger, 1975), chapter 6.

$400 to $1200 per pupil in different sites, with the average around $800.[3] A CPIR survey of Follow Through projects in 1971-72 estimated that of $52.4 million expended for Follow Through in that year, 47 percent went for educative services, 30 percent for project support services and 19 percent for pupil services.[4]

Plans drawn up by SRI in late summer 1971 included a proposal for collecting Follow Through cost data, but the idea was rejected by OE reviewers, in part because "the objectives of the cost study and the major issues which will be explored in this study" were not made explicit and because "SRI did not indicate whether the CPIR form includes measures which are adequate for answering the specific type of questions to which the Follow Through evaluation is addressed."[5] In 1973, the OE Follow Through Evaluation Review Panel strongly recommended that the evaluation include "an accurate assessment of the total costs (from which sources) spent on both Follow Through and comparison children," but gave no details on what sort of cost study ought be done.[6] No further action seems to have been taken on the idea of a cost analysis of Follow Through until 1975, when OE issued an RFP for a study of Follow Through costs. The RFP stated that

---

[3] Emrick, Interim Evaluation, pp. 100-270. FT per-pupil expenditures were calculated as "total anticipated funds supplementing district maintenance of effort (sometimes includes Title I funding) divided by project enrollment" (p. 87

[4] Ronald N. Jessee, "The Follow Through Program," National Center for Educational Statistics, Consolidated Program Information Report. NCES-75-309D. Washington, D.C.: NCES, 1975. Only 95% of the 150 agencies surveyed in this study responded. Data in the survey report are based strictly on these reports. Data from them indicate that per-pupil program costs in 1971-72 were approximately $750-$760.

[5] J. Crooks and D. Burnes, memorandum "Consolidated Review of September 17 Deliverables" to Mr. Fred Bresnick, Acting Chief Research and Evaluation Section Follow Through Program, October 6, 1971.

[6] OE, Second Report of the Follow Through Evaluation Review Panel, March 1973, p. 12. Parentheses in original.

> [C]ost information on Follow Through models is needed for two purposes: (1) to help OE make judgments about which model to promote for widespread replication, and (2) to provide local education agencies that might be considering adoption of a Follow Through model with the necessary data on costs.[7]

RMC Research Corporation won the contract competition and in the summer of 1975 began work on the cost analysis of Follow Through projects. RMC researchers used what they called a resource approach to cost analysis. They identified around 100 different resources used in Follow Through projects and comparison classes, and then via written questionnaires and interviews, derived estimates of how much of each resource was used in selected samples of Follow Through and comparison classrooms. Informants on resource use were school principals and teachers, and for Follow Through classes, the local Follow Through project directors. After the "quantity of each resource available to the classroom" was determined, "that quantity was multiplied by a standard cost to determine the cost per class for each resource. The per class costs were then divided by the number of children in the class to yield an estimate of the average cost per child for each resource."[8] Cost estimates were aggregated to produce overall estimates of costs per class of the resources used in Follow Through and Non-Follow Through classes. Also, average costs per class for each of the Follow Through models included in the study were estimated, and incremental costs of Follow Through were calculated as "the average of the differences between FT and NFT per child expenditures for each model."[9]

---

[7] OE, RFP 75-67, as quoted in RMC Proposal Cost Analysis Study of Follow Through Projects (Bethesda, MD: RMC Research Corporation, April 18, 1975), p. 3. The RMC Cost Analysis report reflects mainly the first of the two purposes mentioned above. Another documents currently being prepared is aimed to serve as a source of information for LEA's that might be considering adoption of a Follow Through model.

[8] Donald Rogers, et al., A Cost Analysis of Follow Through Projects (Bethesda, MD: RMC, April 1977), p. xii. The RMC cost analysis was used as volume IV in OE's series The Follow Through Planned Variation Experiment.

The findings of the RMC Cost Analysis report are summarized briefly in Table 15. Several weaknesses in the cost study can be mentioned. First of all, the validity of the resource information obtained from school principals and Follow Through project directors was never ascertained in terms of observations of resources actually used in Follow Through and comparison classes. Second, the resource approach was not used uniformly throughout the RMC study. Sponsor costs were estimated not in terms of actual resources required or provided by sponsors but instead solely on the basis of budgets in sponsor applications to USOE--even though the validity of budgets as sources of cost information had been denigrated elsewhere in the RMC report.[1] Nevertheless, these weaknesses seem minor compared to two flaws in the RMC cost study which led to several "findings" which seem patently non-sensical.

The major problem with the RMC study is that it never comes to grips with the issue of what a Follow Through model is. The RMC researchers took the totally agnostic approach of pretending that a Follow Through model was whatever there was in a classroom ostensibly affiliated with a particular model.

---

[9] For a listing of the sample of projects and models in the RMC study see Tables 4 and 5 on pp. 102 and 103 respectively of this report.

[1] Compare for example pp. 46-7 with p. 3 in the Cost Analysis.

Table 15: Summary of the Findings from the RMC Cost Analysis of Follow Through Projects

"The study results show statistically significant differences in resource utilization between Follow Through and comparison programs. The average Follow Through class costs $19,000 more than the average comparison class. The average cost per Follow Through child is almost $800 greater than the average cost per child in comparison programs." The increased expenditures for Follow Through were distributed, overall on the average, as follows:

Teacher Aides	26%
Follow Through Sponsor Costs	21%
Training, Medical, Psychological, Social and Other Services	16%
Project Director and LEA Evaluators	9%
Curriculum Developers	9%
Parent Coordinator/Home Visitor	6%
Other Costs	13%

"The study did not uncover any statistically significant differences in total costs among the Follow Through models. However, the average total cost of operating a Follow Through model did vary considerably from model to model. Furthermore, the cost of operating a specific model varied considerably from LEA to LEA. Therefore, although large differences in total costs were observed among Follow Through models, the interpretation of those differences was strongly affected by the amount of variation across LEAs associated with the estimated total cost of each model. This cross-LEA variation was great enough to prevent any finding of statistically significant differences in total costs."

Observed costs per class ranged from roughly $24 to $44 thousand in Non-Follow Through schools and from approximately $31 to $55 thousand in Follow Through schools. These data "graphically reflect the high variability of costs across LEAs and models. Some of this variability can be explained by the relationship between the costs of Follow Through and non-Follow Through schools in the same LEA. As the cost of operating non-Follow Through schools increases, the cost of operating Follow Through models also increases--regardless of the model being operated. This relationship suggests that the observed costs are biased for models in LEAs with either primarily high- or primarily low-cost non-Follow Through programs. To compensate for this bias, adjusted total costs were formulated."

"...The adjusted cost was formulated by calculating an incremental cost for each model and adding it to national average per pupil expenditures. The incremental cost consisted of the average of the differences between Follow Through and non-Follow Through per child expenditures observed for each model. The national average per child expenditure was found to be $1,282. Because the incremental cost was considered the best estimate of the increase in per child expenditure associated with the operation of a Follow Through model and because the national average was considered the best estimate of the per child expenditure in an average LEA, the sum of a model's incremental cost and the national average was considered the best estimate of the cost per child of operating a specific Follow Through model in an average LEA."

Incremental model costs per pupil ranged from a low of around $100 for one model to a high of $1400 for another, with the median for all models included in the study around $790. When added to the estimated national average per child expenditure of $1,282 these incremental costs yielded "adjusted estimated costs of operation" per child ranging from a low of $1,380 for one model to a high of $2,700 for another model.

Source: Rogers, et al., Cost Analysis, pp. v-ix, and p. 73. The RMC summary of the cost analysis report contains two figures and a table which are not included, but instead information from them only briefly described in this summary.

Hence the RMC study found that the average cost per class for operating a particular model could vary by thousands of dollars; in the extreme case, it was found that the BOPTA model of Georgia State University was operated at a per-class cost ranging from $31,000 to $55,000.[2] Moreover, since many LEA's use more resources in their comparison programs than many LEA's use to operate Follow Through, the RMC analysts suggested that "some LEA's with programs costing above average might operate a Follow Through model with the same amount of resources they are currently using on less."[3] Such a conclusion means something only if we know what is meant by "operating a Follow Through model." Yet the RMC researchers never attempt to define explicitly what this means. Operationally is seems that they defined a "model" only in terms of the resources reportedly used in a project affiliated with a Follow Through model. In these terms an LEA clearly could "operate" a Follow Through model for less than they currently invest--all they would have to do is use fewer resources and call their program a model!

The second major flaw in the Cost Analysis is the manner in which the RMC researchers calculated "incremental costs" of operating a Follow Through model. "Incremental cost was calculated as the difference in total costs between the schools that have implemented the FT project and comparison schools in the same LEA" (pp. 22-3). The problem was that RMC seems to have had considerable difficulty in locating appropriate comparison schools. "The actual selection of the comparison school was left to the Follow Through

---

[2] See Figure S-2 in Rogers, et al., Cost Analysis, p. 83.

[3] Rogers, et al., Cost Analysis, p. 83.

Director because non-Follow Through schools were generally unwilling to participate in the survey" (p. 76). In fact, in two LEA's out of around 50 included in the study, "all the potential comparison schools refused to participate in the study" (p. 77). Moreover, even when comparison schools could be coaxed into participating, "the programs in these schools were not documented. Therefore, the comparison sample can only be described in a very general way. The most accurate description appears to be that the comparison schools provide compensatory education through ESEA Title I programs to children with socioeconomic backgrounds similar to Follow Through children."[4]

Nevertheless the RMC data make it clear that in some cases the comparison groups did not provide very good estimates of what resources Follow Through schools would have provided had they not been operating a Follow Through model. For seven Follow Through models, incremental costs in at least one of eight resource categories were negative--implying that schools would have provided more resources had they not been operating a Follow Through model. As the authors of the RMC report themselves point out:

> The incremental cost approach attempts to net out the LEA effect by assuming the resources avilable to a Follow Through school would equal the resources available to the comparison school in the absence of Follow Through. However, that assumption is questionable. The comparison school used to derive the incremental cost may be atypical for the district. Even if the LEA effect could be removed, the model/model comparisons might still be influenced by other factors. For example, the extent to which Office of Education policies imposed resource structures on the models would mitigate the extent

---

[4] Rogers, et al., Cost Analysis, p. 77. The comparison schools used in the RMC cost study were generally not the ones used in the national evaluation of children's test data. According to Appendix A of the RMC report only around one-quarter of the RMC comparison schools were also included in the national evaluation.

> to which the comparison could be interpreted as analyzing the differences in the amount of resources used to operate a model. For example, in making grants to Follow Through site, OE may have mistakenly provided too much or not enough funds to correctly operate a model. (p. 90)

So in the end the RMC Cost Analysis seems to have foundered on two of the familiar problems which have so long bedeviled the national evaluation. After ten years and tens of millions of dollars, no one has yet come up with a clear-cut method of determining the extent to which a project is operating--or implementing--a particular Follow Through model. And in the cost study as in the national evaluation of children's test scores, the use of comparison groups has not proved to be a very effective method for estimating what things would have been like had there been no Follow Through program or models.

Accomplishments of the National Follow Through Evaluation. The Follow Through evaluation is not only an evaluation of alternative approaches to early education. As Richard Anderson, director of the second phase of the Abt evaluation effort, pointed out recently, it is also an evaluation of the planned variation approach to evaluation.[5] It is clearly too early for any final assessment of the Follow Through experiment in evaluation, but even now several points seem clear.

Perhaps most obvious is that judgments on the worth of the Follow Through planned variation evaluation range widely depending on the standard of comparison one uses. Compared to earlier national evaluations of compensatory education like the Westinghouse-Ohio evaluation of Head Start or the national evaluations of Title I, the Follow Through evaluation does not look too bad. The Follow Through evaluation was an improvement over the Westinghouse-Ohio study in that it employed pre-tests as well as post-tests. Also the Follow Through evaluation seems to be a clear improvement over several early evaluations of the Title I compensatory education effort. The 1967 TEMPO study of Title I, for example, was judged a "complete washout" which "did not show anything at all one way or another."[6] The successor to the TEMPO evaluation, the Belmont Joint Federal/State Coordinated Evaluation Effort was similarly disastrous. Though costing millions the Belmont system failed to result even in any useable data.[7]

---

[5] Richard B. Anderson, "Follow Through: Testing One Model of Evaluation and Several Models of Compensation," paper presented at the annual meeting of the American Educational Research Association, San Francisco, 20 April 1976.

[6] Milbrey W. McLaughlin, Evaluation and Reform The Elementary and Secondary Education Act of 1965/Title I. (Cambridge, MA: Ballinger, 1975) p. 37. "Complete washout" is McLaughlin's description of the TEMPO study. "Did not show anything one way or another" are Alice Rivlin's words describing the findings of the TEMPO study.

[7] McLaughlin, Evaluation and Reform, p. 105.

Compared even to its own history the national Follow Through Evaluation looks not too bad. The Emrick report was improved over the Pittsburgh evaluation if for no other reason than because the SRI investigators included post-test scores as well as pre-test scores in their analysis. The reports resulting from the first phase at Abt can be viewed as improved over the Emrick report because they used more standard outcome measures instead of relying upon an overlapping collection of measures developed ad hoc during the course of the evaluation. And the Abt III and Abt IV reports can be seen as improved over the first two Abt studies because alternative explanations of results were more fully explored (the preschool confounding issue, and the covariate unreliability problems, for example) and analyses were more fully reported.

From such an account the national evaluation might seem to be a model of progress. From another perspective a very different picture emerges. For if the evaluation has improved over time, it has been getting better at answering narrower and narrower questions. The Pittsburgh evaluation set out to examine program processes, outcomes and costs. The SRI evaluation had similarly broad aspirations in early years but was gradually narrowed to look mainly at outcomes--though with some attention to processes, which developed into Jane Stallings' classroom observation studies. When the evaluation analyses shifted to Abt, several sorts of questions, regarding institutional change, for example, were abandoned. And as the Abt analyses moved toward completion several more questions were sloughed off--most notably regarding model implementation. At the end, other issues that did remain, like the analyses of the parent interview and teacher questionnaire data, were so abstract and far removed from the questions that motivated them, that very few if any answers were evident.

So in the end the national Follow Through evaluation came to address only a few of the evaluation criteria suggested as appropriate to the program, and to focus on effects questions, with relatively little regard for descriptive, process or cost issues. The treatment-control paradigm was the main model shaping the inquiry into program effects, but there have been many issues troubling the application of this paradigm to the case of Follow Through. Treatments are only dimly perceived. Definitions of key elements in the paradigm, like the treatment group and treatment subjects, never have been entirely clear. The appropriateness of control or comparison groups has remained an uncertain issue throughout the evaluation. Also, the four-year time span between pre- and post-tests led to an attrition problem of considerable proportions--around 70% of the children pretested in fall kindergarten in both cohorts II and III were not around to be tested in spring of third grade. Collectively, these problems have made it difficult to draw many clear-cut findings from the national evaluation. The one unambiguous finding seems to be that basic skills sponsors and in particular the Direct Instruction approach can help raise children's basic skills test scores.

But if substantive findings are scarce, the Follow Through evaluation experience does seem to offer a number of lessons about evaluation. Usually these are couched in terms of the treatment-control experimental paradigm. For example, the GAO review, Follow Through: Lessons Learned from Its Evaluation and Opportunities to Improve Its Administration, concluded with a recommendation that the Secretary of HEW

> --Develop greater safeguards to insure that future experimental programs are not designed apart from evaluation to maximize the degree to which experimental results will be statistically reliable (p. 26).

Specifically, the GAO review mentioned the "lack of random assignment" in Follow Through and "the lack of comparable control groups." Similar lessons were drawn from Richard Elmore's review of Follow Through as a large-scale social experiment. Elmore focused particularly on the problems of treatment specification and implementation and on effects measurement. Lessons drawn from the Follow Through evaluation by GAO and by Elmore clearly warrant attention--especially if anyone sets out again to carry out a large-scale planned variation experiment. In this respect Elmore's caution on the need to limit the scale of such an endeavor until treatment programs can be developed and well specified seems especially important.[8] His comments on the point coincide closely with recommendations by David Weikart and Bernard Banet on "Model Design Problems in Follow Through."[9]

The difficulty with the sorts of lessons learned is that they take for granted the planned variation conception of Follow Through--a conception which is clearly not shared by many of the participants in Follow Through. Also these sorts of lessons are not terribly new. As Elmore points out, Follow Through administrators recognized the problems of specification, measurement and implementation from the onset of the evaluation.[1] Numerous attempts have been made to specify Follow Through treatments; tremendous effort has gone into developing new measures for use in the Follow Through evaluation; and the implementation issue has been worried over and attacked repeatedly; but all to little avail.

---

[8] Elmore, Follow Through: Decisionmaking, p. 377.

[9] Rivlin and Timpane, Planned Variation in Education, pp. 61-77.

[1] Elmore, Follow Through: Decisionmaking, pp. 375-377.

The lack of progress on these issues might be attributed to bad management or to a shortage of expertise. As the GAO and HEW Audit Agency reviews of the Follow Through evaluation suggest, there may have been considerable room for improvement in the program management domain. But at the same time, this sort of explanation for the problems in the evaluation rings somewhat hollow. Key people in the Follow Through evaluation story, like Robert Egbert and Garry McDaniels, have after all, been widely recognized for their work in Follow Through and have advanced to other responsible administrative positions. And the list of experts who have consulted on the Follow Through program and evaluation at one time or another reads like a Who's Who in educational research and evaluation--Kuno Beller, Urie Bronfenbrenner, Donald Campbell, David Cohen, Jacob Cohen, Leon Festinger, Eleanor Maccoby, Alexander Mood, Richard Light, Barak Rosenshine, Sheldon White, to name just a few. Likely there is some cause for questiniong the role of these big-name consultants or "shadow experts" as Guttman and Willner might call them. At the same time, however, their participation suggests that the problems troubling the evaluation were not simply those of technical expertise. For some this directs attention to organizational issues. Richard Elmore, for example, drew a number of "organizational lessons" from his review of Follow Through as a large-scale social experiment:

> --Organization-building is a basic feature of large-scale social experiments; designs have no meaning until they are translated into an organizational mechanism for carrying them out. It was not the preoccupation of Follow Through administrators with organization building that created problems for the experiment, but their failure to design an organization around the purpose of experimentation.

> --In the conduct of large-scale social experiments, program decisions are design decisions and vice versa. The failure of Follow Through administrators to recognize that they were making design decisions in the construction of the program resulted in an organization of such scale and complexity that it couldn't be translated into a coherent design.
>
> --The intervention of outsiders in the Follow Through evaluation proved to be a singularly ineffective way of improving it. In fact, most outsiders did more to confuse than to clarify the basic problem--lack of a coherent design. Significant improvements in the evaluation came only after it was brought under control from the inside.[2]

Elmore's "organizational lessons" are intriguing. Doubtless there is room for improvement in the organization of programs like Follow Through and evaluations of them, but another interpretation is that organizational problems arose precisely because Follow Through never was a planned variation experiment to many of the people engaged in it. Indeed, from this point of view, the Follow Through evaluation experience has to give us pause to question the appropriateness of the experimental paradigm that has come to predominate among evaluators of Follow Through.

The general popularity of the experimental paradigm among evaluators is part of an effort over the last decade or so to make program evaluations more rigorous. "Evaluation studies are generally of low quality," observed Edward Suchman in 1967, but "By adopting the scientific method, the hope is that the results of the evaluation study will be more objective and of ascertainable reliability and validity."[3] The trend has been reflected in the growing popularity of terms like "evaluative research" and "evaluation research." These hybrid terms reflect the notion that evaluations can be improved by making them more like research--and probably also the hope that the aura of research will rub off and brighten the status of evaluators. But the trend, and the hybrid terms sprouting from it, obscure a fundamental difference between evaluation and research. In its root sense evaluation means to determine the worth or value of something. Program evaluation usually refers to the determination of the worth or value of a program in terms of program goals in order to inform decision-making about it. Research generally implies a more agnostic inquiry to establish facts or principles of some sort. Research generally is little attuned to issues of values or to decision-making processes.

---

[2] Richard Elmore, Follow Through: Decisionmaking, p. 390.

[3] Edward A. Suchman, Evaluative Research (New York: Russell Sage, 1967), p. 20

The Follow Through evaluation, it seems to me, can be viewed as a sort of case-study of the weaknesses of the recent fashion of evaluation as experimental research. Research-oriented evaluators tend to become entranced with analytical methodology and are little concerned with program description. Evaluation researchers, at least in Follow Through, invariably seem to have conducted far more analyses than ever show up in final reports, and reports tend to be far more heavily oriented to continuing evaluation research than to program decision-making. Generally reports are also extremely lengthy and are not at all easy to read. The major Follow Through evaluation reports, for example, average around two pounds in weight and are so complex that days of concentrated effort are needed to plow through any one of them. Moreover, major evaluation reports have usually been very slow to reach final form. The time lag between last data-gathering and publication of final report has generally been around two years. Even the unanalyzed data in the SRI feedback tables took eight months or more to prepare and send back to Follow Through projects. Such lags, coupled with the complexity of major Follow Through evaluation studies, make their use for program decision-making highly problematical. These problems could be elaborated at some length, but instead let me turn to what may be the biggest drawback to the research orientation of Follow Through evaluation efforts.

Complex social programs like Project Follow Through have a variety of interlocking and not terribly well-specified goals, usually reflecting a variety of values and compromises among interests. Conflict over the main meaning of Follow Through has a long history, resulting in 1974 in that peculiar official description of Follow Through as an "experimental community services program designed to assist, in a research setting, the overall development of children..." The official description was an uncomfortable compromise among interests holding different views of the program and these views continue even now to emphasize different aspects of the program. The social service view of Follow Through highlights the extra services--both instructional and noninstructional--provided by the program. A social action orientation tends to accent parental and community involvement in Follow Through. A research and development slant tends to focus on the support Follow Through has given to the development of alternative curriculum approaches. The planned variation concept to which the national evaluation has been latched since 1972 gives predominance to the comparative assessment of the effects of the alternative approaches. While there was a variety reasons for latching the Follow Through evaluation onto this particular visio of Follow Through, it seems clear that one reason for this link-up was the research aspirations of evaluators. After all, the planned variation notion itself was a product of the search for more scientific and objective means for evaluation.

Fitting the program to the Procrustean paradigm of treatments and control groups, and assumptions of equivalence, has clearly had a price. Aspects of the program which did not fit the paradigm--like parental involvement and comprehensive services--have been largely ignored in major national evaluation reports. Even within the planned variation conception, the evaluation became quite narrow--reduced to only four standard outcome measures. The 1972-73 review panel acknowledged that the limited reliance of the evaluation on only those four outcomes would leave "unanswered many questions about children's learning ability," but the technical inadequacies of other test instruments led the panel to recommend only the MAT, the Ravens, the IARS and the Coopersmith for the final assessment at the end of third grade. This is resulting, in one sponsor's words, in an assessment which "stand[s] up as poor, puny and almost perverse shadows of Follow Through reality."

So it did not quite work as John Hughes had envisioned in 1968. Instead of bringing "the knowledge and technical expertise of the educational specialist, the systems engineer, and the behavioral scientist...into harmony with the pluralistic value structure of our society," the planned variation evaluation seems to have turned Hughes' hope on its head. The values reflected in the different visions of Follow Through have been whittled to fit the technology of the experts. So the fit between the program and the planned variation evaluation has not been very good. And since the program was not originally designed as a planned variation experiment, even the part of the program that could be squeezed into the treatment-control paradigm of experimental research did not fit very well.

From the view of Follow Through as a planned variation experiment, the result may be disappointing. Yet it is useful to recognize that many of the weaknesses of Follow Through as a planned variation experiment may be viewed quite differently in terms of other conceptions of the program. For example, the fact that Follow Through "control groups" generally were more advantaged than Follow Through treatment groups at the start of the program has been continually troublesome for evaluators. From the service delivery point of view, this can be seen as evidence that Follow Through has been at least somewhat successful in reaching the disadvantaged population it was intended to serve. The stubborn problem of treatment specification seems to have resulted, at least in part, from Follow Through sponsors' continuing development and adaptation of their curriculum approaches. And the very continuation of the program, despite periodic Office of Education efforts to phase it out, might be viewed as evidence of success for one of the social action goals of Follow Through, namely an increased locus-of-control for program participants.

These interpretations are surely over simple, but they point up the fact that research-oriented evaluation--aiming at estimation of program effects on specific test outcome measures--is capable of illuminating only part of the value of a complex social program like Follow Through. The results of the national Follow Through evaluation in terms of children's third grade test scores seem disappointing--when contrasted with comparison Title I programs, only one or two Follow Through models seem to have measurably boosted children's third grade test performance. Yet Follow Through remains a highly popular program--as evidenced in Congressional continuation of the program despite repeated OE efforts to phase it out. In part the differential valuation of Follow Through seems to derive from consideration of different parts of the overall Follow Through program. The national evaluation has come to focus almost exclusively on children's test scores, while members of Congress seem to be responding, as a result of communications from their constituents, to broader aspects of the program, including medical and dental services and parental involvement.

More fundamentally, however, the differential valuation of Follow Through seems to derive from different kinds of evidence about the program. Research-oriented evaluators tend to give most credence to experimental evidence and "hard" data, and relegate hearsay and opinion to small status. Congressmen and other decisionmakers, however, tend to give greater weight to the opinions of their constituents and give lesser weight to research evidence, especially when, as in the case of the Follow Through evaluation, even presumed experts cannot agree on the validity of many of the research "findings." The main lesson I draw from the Follow Through experience is that the burden of responsibility for reconciling disparate valuations of programs such as Follow Through lies not with program participants, but with evaluators. Experimental evidence concerning program effects cannot, after all, tell us anything about the social meaning of such effects or or the programs that produced them. If evaluation in its root sense is to be accomplished as an effective aid to decision-making, evaluators must bridge the gap between experimental evidence of programs' effectiveness and other valuations of their social meaning.

Glossary of Acronyms

ANCOVA - Analysis of Covariance

BESE - Bureau of Elementary and Secondary Education

BRACE - Behavioral Ratings and Analysis of Communication in Education

EDC - Education Development Center

EF - Entering first grade

EK - Entering kindergarten

ERIC - Educational Resources Information Center

FK - Fall kindergarten

FT - Follow Through

GAO - Government Accounting Office

GPO - Government Printing Office

HEW - Health, Education and Welfare

HSPV - Head Start Planned Variation

IARS - Intellectual Achievement Responsibility Scale

LEA - Local Education Agency

MAT - Metropolitan Achievement Test

NFT - Non-Follow Through

NIE - National Institute of Education

OCD - Office of Child Development

OE or USOE - United States Office of Education

OEO - Office of Economic Opportunity

OPBE - Office of Planning, Budgeting and Evaluation

PAC - Policy Advisory Committee (sometimes also deciphered to mean Parent Advisory Committee, since according to FT rules and regulations, "More than one-half of Policy Advisory Committee members shall be low income FT parents..." In a couple of instances, this acronym has also been deciphered as Political Action Committee).

PEES - Pacific Educational Evaluation Systems

PL - Public Law

PPVT - Peabody Picture Vocabulary Test

PSI - Caldwell Preschool Inventory

R&D - Research and development

RFP - Request for Proposals

RMC - Resource Management Corporation

SEA - State Education Agency

SK - Spring kindergarten

SRI - Stanford Research Institute

S1 - Spring first grade

S2 - Spring second grade

S3 - Spring third grade

UA - Unit of analysis

WRAT - Wide Range Achievement Test

	1967	1968
National Events	Jan. - President Johnson in State of Union Address advocates a follow through on Head Start. Feb. - Johnson proposes Follow Through to Congress. Mar. - Johnson in poverty message to Congress proposes $110 M for "HS follow through program." Dec. - Congress approves EOA amendment, and FT, but slashes budget for OEO.	Feb. - LBJ proclaims fifth freedon--from ignorance--and asks Congress to expand HS and FT. Calls for rise in $ for HS from $340 M to $380 M of which $36 M is for FT.
Program History	May - Selected nominated communities invited to submit applications to participate in FT. Jun. - 30 pilot sites announced. Aug. - Egbert chosen as Follow Through director. Sep. - Follow Through programs begin in 30 sites. Late summer - In response to "pressure from big-city mayors and congressional delegations" ten additional pilot sites selected to begin second semester of '67-'68 school year.	Jan. - Meeting in Washington for prospective Sponsors. Feb. - Kansas City meetings to pair projects with Sponsors. Oct. - At Atlanta meeting general consultants charge sponsors and evaluators with slighting parent, community and minority involvement in FT.
Evaluation History	Aug. - Evaluation contract awarded to University of Pittsburgh.	Jun. - Evaluation contract signed with Stanford Research Institute. Aug. - Evaluation planning meeting at Menlo Park raises basic issues about evaluation; most left unresolved.
Major Evaluation Reports		

1969	1970	1971
Jan. - In budget message, LBJ requests increase in HS-FT budget. - President Nixon inaugurated. Feb. - Nixon proposes expansion in education programs, transfer of HS to HEW. Alluding to Westinghouse evaluation of HS, notes long-term effects of HS appear small. Mar. - OCD created. Dec. - Senate-House conference committee agrees on 2-[illegible] extension of OEO.	Jan. - Nixon vetoes appropriations bill including antipoverty funds. Mar. - In special message, Nixon holds compensatory education has not measurably helped disadvantaged children and should not be expanded pending further study. Aug. - Nixon vetoes $4.4 B education bill as inflationary. Oct. - Nixon Administration asks OEO to cut $150 M from budget, including HS funding cut.	Mar. - Nixon Administration sends to Congress bill to renew and antipoverty legislation for two years. Sep. - Senate passes comprehensive new child development bi including expansion of Head Start. Dec. - House gives final Congressional approval to antipove child development bill. - Nixon vetoes combined EOA/Child Development bill.
Feb. - Draft Follow Through Program Manual. May - Memo of understanding redelegating authority for FT from OEO to HEW.	Nov. - Egbert announces intention to resign.	Jul. - Egbert leaves and Rosemary Wilson is named acting director of Follow Through. Oct. - First FT Sponsors Conference recommends study of FT experiment in developing educational alternatives. Dec. - Regional meetings held to review criteria by which projects to be judged for continued funding.
Apr. - Pajaro Dunes conference. SRI evaluation plans criticized as unfair to some sponsors. Jul. - & Aug. SRI criticized for focusing overmuch on traditional measures. Nov. - After more criticisms SRI reorganizes evaluation effort.	Feb. - In review of preliminary findings, evaluation criticized as "mindless empiricism." May - Goldberg review panel recommends SRI continue evaluation, but asks for clarification of objectives, schedules, etc. Jun. - SRI proposal for third year of evaluation approved. Oct. - & Nov. Preliminary SRI reports on 1969-70 data incite controversy.	Feb. - R. Fairley visits SRI, asks for identification of b sponsors in grade-equivalent terms. Mar. - PEES report criticizes management of SRI contract, understaffing of SRI project. Jun. - SRI proposal for contract continuation approved, wi Don Rose of PPE/BESE as project officer. Jul. - At Menlo Park conf., Fairley discusses 5-year plan expansion of FT & criteria for "Consumer's Guide." Aug. - GAO report cites SRI contract as example of need fo more specific contract objectives, and criticizes Pittsburgh evaluation. - Snyder leaves position as chief of FT evaluation. Sep. - SRI submits planning documents, wide reviews "highl critical." Nov. - McDaniels appointed acting chief of FT evaluation. Dec. - Huron Institute gets contract to assist in FT evaluation planning. - At Washington meeting, decision to have SRI produce progress report instead of Consumer's Guide.
Apr. - Westinghouse-Ohio HS Evaluation released, leading to much controversy. May - Kurfoerst & Stephens, Univ. of Pittsburgh, Follow Through Evaluation for 1967-68. Dec. - SRI, "Longitudinal Eval. of the National FT Program, July '68-November '69," including appendices A-Q.	Sep. - SRI, Appendix H. Program Description, Vols. I-XI. Dec. - SRI, Appendix D. Issues in Noncognitive Measurement.	Mar. - SRI, Draft, Longitudinal Evaluation of Selected Features 1969-70, plus Appendices A & B. - Fox, An Examination of Support Systems...

1972	1973
Jan. - Ed. Comm. Marland supports cutting comp. ed. programs like ESEA & HS which have "not improved quality" of education for disadvantaged. Mar. - Nixon message to Congress calls for Ed. Opp. Act to be funded from ESEA plus additional $. Apr. - Senate subcom. approves bill extending EOA with concessions to Nixon administration. May - In supplemental appropriation, Senate subcomm. allocates additional $3 M to FT, giving new lease on life to 21 of 26 projects which OE had slated for termination in 72-73 school year. Aug. - Nixon vetoes $30 B Labor-HEW appropriation bill. Sep. - Congress approves and Nixon signs 2-year extension of EOA authorizing $70 M for FT for FY's '73 and '74.	Jan. - Pres. Nixon inaugurated for second term. - Nixon budget proposes dissolution of OEO & cuts or phasing-out of several educ. programs including HS & FT. - Admin. plans for educ. revenue sharing disclosed. Mar. - Nixon in State of Union message on human resources says elimination of OEO is organizational rather than substantive change. Apr. - Fed. Judge brands efforts to dismantle OEO illegal, & orders halt to curtailment of antipoverty program in FY'73. Jul. - All but 3 OEO programs to be transferred to other departments.
Jan. - Mar. - FT project-monitoring system reorganized into five area-teams. Feb. - R. Fairley, Director of Comp. Educ., announces 5-year plan for FT, including diffusion of FT successes to Title I programs, Consumer's Guide to FT, involvement of SEA's and gradual phaseout of FT. - Federal Register announces grants to SEA's for development of successful FT program models. Apr. - OE announces termination of 26 FT projects. - Congressmen decry abrupt termination of FT projects. Jun. - FT planning grants made to 6 SEA's. Funding restored to 21 FT projects. Aug. - Rosemary Wilson named full director of FT program. Oct. - At Denver national FT conf., parents decry exclusion of parents from FT eval., evaluation's narrow focus, & OE plans for phase-out of FT.	Aug. - Projects officially notified that "experimental" FT program will begin phasing out in the 1974-75 year at rate of one grade-level per year.
Jan. - SRI submits materials for OE report to Congress. Mar. - RFP for FT evaluation analysis. May - Guttman-Willner letter to Richardson charges abuse in FT evaluation contract. Jun. - Evaluation contract awarded to Abt. SRI retains data collection contract. Jul. - McDaniels asks sponsors to make objectives explicit & plan for evaluating those not in national eval. Fall - R. Marciano becomes SRI project director. Dec. - HEW Audit Agency report criticizes management of evaluation contract. - OE issues RFP for FT data collection.	Mar. - Report of review panel recommends comprehensive and integrated evaluation of FT. Apr. - OE issues RFP for technical assistance on FT evaluation. - McDaniels leaves as chief of FT evaluation. - Eugene Tucker becomes chief of FT evaluation and evaluation gradually moved to OPBE.
Jan. - David, "Rationale for Selection of Sites...for Spring 1972 Testing." Feb. - SRI produces "Preliminary Guide to FT" as one of progress reports. Aug. - SRI, Draft Evaluation of National FT Program 1969-71, including annexes I-IV and appendices A-F. Dec. - HEW Audit Agency, review of SRI contracts.	Jan. - Abt, "Interim Report: An Analysis of Selected Data, 1969-72." Feb. - Emrich, Interim Evaluation 1969-71. Mar. - OE, "Second Report of FT Evaluation Review Panel." Apr. - Krulee, Organizational Analysis. Jun. - Soar, FT Classroom Process Measurement & Pupil Growth (1970-71). Aug. - Stallings, FT Program Classroom Observation Evaluation 1971-72.

1974	1975
Feb. - Nixon budget for FY '75 includes plans for FT phase-out; FT budget request $35 M. May - House approves allocation of $1.5 B to HS & $180 M to FT for FY '75-77, but votes to abolish OEO, transfer HS & FT to HEW. - In supplemental appropriations, Congress provides additional $12 M for FT. Aug. - In first news conf., Pres. Ford indicates his view on antipoverty programs same as Nixon's, no plans to revive OEO. Nov - HR Comm. on Labor & Public Welfare report "specifically rejects any suggestion that FT has outlived its usefulness and should be phased out." Dec. - Senate approves compromise bill extending life of OEO but changing its name to Community Services Admin. Jun. - Reversing previous decision OE decides to fund new entering class in FT in fall '74. - Federal Register publishes interim rules and regulations for FT. Oct. - Clash over "horse race" eval. versus "armies of mothers," at annual sponsor conference.	Jan. - Pres. Ford signs HS, Econ. Opp. & Comm. Partnership Act of 1974, authorizes $60 M for FT for FY '75. Mar. - In testimony to House Appropriations subcommittee, OE officials request $41.5 M for FT for FY'76, $5.5 M less than requested for previous FY. Request based on plan to phase out FT. Nov. - Ford asks for cuts in Fed. educ. funds, including $17.5 M cut for FT. Dec. - Appropriation bill that includes $624 M for antipoverty programs, passed by Senate & sent to White House. Apr. - Federal Register publishes final rules and regulations for Follow Through.
Jan. - FT evaluation officially transferred from Bureau of Elementary and Secondary Education to OPBE. Sep. - New RFP for Follow Through evaluation analysis. Dec. - Evaluation contract returned to Abt with R. Anderson as director of project.	Mar. - First planning conference for third-year Abt evaluation analysis, limited mainly to evaluators, issue of how to handle pre-school confounding left open. May - RMC awarded contract to perform cost analysis of FT. Jun. - Second planning conference on Abt evaluation analyses, pre-school confounding issue again unresolved.
Feb. - David, Draft Follow Through Summer Study. Mar. - Cline, Abt I, Vols. A & B. Aug. - Stallings, FT Classroom Observation Evaluation 1972-73.	Apr. - Judd, Project FT: A Description of Sponsor Implementation Processes. May - Cline, Abt II, Vols. A & B. Oct. - Comptroller General, "Follow Through: Lessons Learned from Its Evaluation and Need to Improve Its Administration."

CHRONOLOGY OF FOLLOW THROUGH PROGRAM, EVALUATION AND RELATED EVENTS

1976	1977
Mar. - Pres. Ford sends to Congress plan to consolidate 24 ed. programs in $3.3 billion block grants to states. - In testimony to House Appropriations subcommittee, OE officials propose plan to phase out entering year of FT in 1977. Request appropriation of only $30 M--i.e., recision of $17.5 M from FY 1976 appropriation. Sep. - Ford vetoes as too expensive $56 B appropriations bill for manpower, health educ. and welfare programs. Oct. - Congress overrides veto of $56 B bill.	Jan. - President Carter inaugurated. Mar. - In Hearings on FY'78 appropriations, OE officials request $59 M for FT, $25 M above request in 1976. Aug. - In testimony to House Ed. & Labor Committee, OE official requests one-year extension of experimental FT program instead of three-year extension asked by Comm. Chairman Perkins.
Nov. - Federal Register publishes notice of proposed rules for selecting Follow Through demonstration sites.	Jun. - Federal Register publishes amended rules and regulations for FT including provisions for demonstration grants. Aug. - Applications of selected FT projects reviewed by Joint Dissemination Review Panel.
Mar. - Memo sent to sponsors requesting 15-page write-up on sponsors' models and own evaluations for OE final evaluation report on FT. Apr. - Washington conference for sponsors to respond to draft of Abt III report. Decision reached to delete concluding chapter of report.	Jun. - Abt IV findings summarized in Education Daily and disputed.
Apr. - Abt, Draft Special Interim Report on Main Effects. May - Wright, Spring '75 Parent Interview Cross-tabular findings. - Sutherland, Spring '75 Teaching Staff Questionnaire Cross-tabular Findings. Mar. - NTL, Facilitating Effective Involvement of Parents, Staff... Jun. - Soar, Appendix E from 1973 report, identifying sponsors. - Rath, FT: Story of Educational Change. Jul. - Stebbins, Abt III, Vols. A & B. - Abt, Draft Addendum to Special Interim Report. Nov. - Wright, Revised Spring '75 Parent Interview Cross-tabular Findings.	Apr. - Stebbins, Abt IV. - Flynn, RMC Cost Analysis.